UNITED STATES OF AMERICA

DEPARTMENT OF HEALTH AND HUMAN SERVICES

FOOD AND DRUG ADMINISTRATION

+ + +

CENTER FOR DEVICES AND RADIOLOGICAL HEALTH

+ + +

GEORGIA TECH - TRIBES
MINIMALLY CLINICALLY IMPORTANT DIFFERENCE (MCID): DEFINING OUTCOME METRICS FOR ORTHOPAEDIC DEVICES

+ + +

November 27, 2012
7:45 a.m.

+ + +

FDA White Oak Campus
10903 New Hampshire Avenue
Building 31 Conference Center (Great Room)
Silver Spring, MD 20993

MODERATOR: FAISAL MIRZA, M.D., FRCSC
Orthopedic Joint Devices Branch
Office of Device Evaluation
CDRH/FDA

Free State Reporting, Inc.
1378 Cape Saint Claire Road
Annapolis, MD 21409
(410) 974-0947

GEORGIA TECH

BARBARA D. BOYAN, Ph.D.
Professor of Biomedical Engineering
Georgia Institute of Technology

L. FRANKLIN BOST, M.B.A.
Professor of Biomedical Engineering
Education Director, Translational Research Institute for Biomedical Engineering & Science (TRIBES)
Georgia Institute of Technology

FDA

MARK MELKERSON, M.S.
Director, Division of Orthopedic Devices
Office of Device Evaluation
CDRH/FDA

DANICA MARINAC-DABIC, M.D., Ph.D.
Director, Division of Epidemiology
Office of Surveillance and Biometrics
CDRH/FDA

WEBCAST: HISTORY AND OVERVIEW OF MCID AND PROs

GORDON GUYATT, M.D.
McMaster University

SESSION 1: PATIENT VARIABLES AND PREDICTORS OF OUTCOME

JEFF SLOAN, Ph.D.
Mayo Clinic

AILEEN M. DAVIS, Ph.D.
Toronto Western Research Institute

LAURA L. TOSI, M.D.
Children's National Medical Center

ROBERT CAMPBELL, M.D.
Children's Hospital of Philadelphia

RON HAYS, Ph.D.
UCLA School of Public Health

TRACI LEONG, Ph.D.
Emory University

DEBORAH MOORE, Moderator
VP, Regulatory and Clinical Affairs
Carticept, Inc.

SESSION 2: OUTCOME INSTRUMENTS

MARC HOCHBERG, M.D., M.P.H.
University of Maryland Medical Center

CHARLES DAY, M.D., M.B.A.
Harvard University

RICHARD COUTTS, M.D.
University of California, San Diego

LYNNE JONES, M.D., Moderator
Johns Hopkins University

SESSION 3: INDUSTRY PERSPECTIVE

KATHLEEN WYRWICH, Ph.D.
United Biosource Corporation
Center for Health Outcomes Research

JAMES RYABY, Ph.D.
Ryaby Associates, LLC

GREG BROWN, M.D., Ph.D.
University of Minnesota

JANICE HOGAN, J.D.
Hogan Lovells

DAVID APPLEBY
Smith & Nephew

CHARLES TURKELSON, Ph.D.
Center for Medical Technology Policy

CHARLES TURKELSON, Ph.D., Moderator
Center for Medical Technology Policy

REMARKS AND ADJOURNMENT

MIKE KEITH, M.D.
MetroHealth Medical Center

ALSO PARTICIPATING

HEATHER STONE, M.P.H.
ORISE Fellow

DR. KEVIN TROYER
ORISE Fellow

VIVEK PINTO

LISA CAMERMAN, Ph.D.
FDA

LEANN SPEERING
Wright Medical Technology

DAVID S. JEVSEVAR, M.D.
Intermountain Healthcare

LISA G. SUTER, M.D.
Yale University School of Medicine

INDEX

PAGE

WELCOME AND INTRODUCTION
Barbara Boyan, Ph.D. 7
Franklin Bost, M.B.A. 10
Faisal Mirza, M.D., FRCSC 14

FDA PERSPECTIVE
Mark Melkerson, M.S. 14
Danica Marinac-Dabic, M.D., Ph.D. 18

INTRODUCTION OF WEBCAST & KEYNOTE SPEAKER
Faisal Mirza, M.D., FRCSC 31
Q&A 53

WEBCAST: HISTORY AND OVERVIEW OF MCID AND PROs
Gordon Guyatt, M.D. 33

SESSION 1: PATIENT VARIABLES AND PREDICTORS OF OUTCOME

MCID Methodology and Clinimetrics - Jeff Sloan, Ph.D. 70

Patient Predictors of Outcome - Aileen Davis, Ph.D. 92

Gender Issues - Laura L. Tosi, M.D. 99

Pediatrics - Robert Campbell, M.D. 107

Patient-Reported Physical Function - Ron Hays, Ph.D. 118

Adjusting for Prognostic Variables - Traci Leong, Ph.D. 128

Q&A and Panel Wrap-Up
Moderator: Deborah J. Moore 133

SESSION 2: OUTCOME INSTRUMENTS

Musculoskeletal and Rheumatologic Perspective on
Outcome Instruments - Marc Hochberg, M.D., M.P.H. 148

Upper Extremity Outcome Instruments -
Charles Day, M.D., M.B.A. 159

Arthroplasty Outcome Instruments - Richard Coutts, M.D. 174

Q&A and Panel Wrap-Up
Moderator: Lynne Jones, M.D. 184

INDEX

PAGE

SESSION 3: INDUSTRY PERSPECTIVE

Overview - Kathleen Wyrwich, Ph.D. 200

Small Company Perspective - James Ryaby, Ph.D. 216

Incorporating Outcome Evidence in Practice - Greg Brown, M.D., Ph.D. 223

Role of the Consultant - Janice Hogan, J.D. 235

Industry Perspective on Patient Factors that Affect Outcomes - David Appleby 243

Q&A and Panel Wrap-Up
Moderator: Charles Turkelson, Ph.D. 248

CLINICIAN SCIENTIST PERSPECTIVE & CONCLUDING REMARKS
Mike Keith, M.D. 263

ADJOURNMENT
Mike Keith, M.D. 273

M E E T I N G

(7:45 a.m.)

DR. BOYAN: Good morning. I'm Barbara Boyan, and I'm a Professor of Biomedical Engineering at Georgia Tech. And it's been my honor to work with Faisal Mirza and the crew from Georgia Tech to help put this conference together.

I'm going to take just a couple of seconds here to remind us why we're here, and then I'm going to turn the podium over to my co-organizer from Georgia Tech, Franklin Bost, who is going to describe to you a little bit about our program at Georgia Tech as we get moving along here.

Let's see if I can figure out how to do the forward -- so we're here to discuss what is the minimally clinically important difference. And this is straight off of our website for the conference, but I think it bears repeating. First, we have to remember that this is not a fixed value. It varies depending on a whole number of variables that include things that I've listed here. But what's really important is why do we need to know this number, and why is it that it has become such an important issue for us to resolve and to try to help as a community of people working in the orthopedic industry how we go about figuring it out.

The workshop was originally designed to address the problem from the point-of-view of experts in the field and to help identify all of the issues that might be involved so that we could arrive at some sort of guidance

to give back to industry and back to our clinical investigators to help them design better studies that would result in the development of safe and effective devices for orthopedics.

And what has been the commonly used metric is the patient-reported outcome. And this has gone on for years. It's something that we understand. It has a sound theoretical basis. It's relevant to the patient group. And it's reliable. And I think most of us would agree that it's valid. But this has not been sufficient really in the device industry to address all of the issues that have come up because devices, the way that they are studied in the concept of a randomized prospective clinical trial, they're not randomized, they're not prospective, surgeons can see what they're using, they know what they look like, they know when they're putting in one and not the other. And so we really need to come up with some other way of determining how small a patient population we can identify that would let us get to the result of a safe and effective device but would let us compare these two datasets so going forward we could use something that would be a more relevant measure.

So the workshop is really a workshop. And we've borrowed the mechanism that's used by the American Academy of Orthopaedic Surgeons to address problems of this kind. First, we're going to have short talks by people who are experts in various aspects of the problem, either patient-reported outcomes, the actual concept of a minimal clinically important difference, and

orthopedic devices themselves. And we've asked people who are academic experts, people who are clinical experts and people who are industry experts because there is an industry side to this, which is the need to find the lowest cost possible to get a device to market that is safe and effective, but so that we can, in fact, invent and develop and manufacture and sell devices for Americans that are going to be used in a reasonable period of time.

So that's Day One. The talks are short. They're going to be rigidly maintained.

All right. Now that we've had that little talk, we're going to get done on time. We're going to get all of the ideas out there with plenty of time -- I hope we've designed in plenty of time for discussion. And then on Day Two, we will come back -- and this is something that is a really special opportunity for us. We're going to break up into clinical groups now, where all of us have had the same body of information, general information, but in each of the clinical specialties, we are hoping that there will be a spectrum of expertise, and we will start to identify the most important issues that we can bring together in a conversation that we can share with FDA so that they can go back and think through what they know that they have to know in order to make a determination about a device, and that we also will have the opportunity to influence each other in the discussion. And I think all of us, whether we're in academics and clinical practice or in industry, realize that we don't have the whole picture and this is an opportunity to get it.

I want to be sure I do this before the groups break up and to say special thanks. This meeting would not have been possible without the American Academy of Orthopaedic Surgeons, which has come through like champs and has funded much of the travel of the orthopedic surgeons who will be presenting to you. I also want to thank the Orthopedic Surgery Manufacturers Association, OSMA, because they stepped in as well, and one small company, amazingly enough, Titan Spine, and we're not even going to be talking about spine products directly, but they felt that this was an important topic to address.

Our organizing team has been fantastic. Martha Willis is the person that most of you have done business with, and also thanks to Maribel Baker, Jenilee Shanks, Jenny Taylor and the Georgia Tech Professional Education Group. There are also people from FDA that have been involved and certainly my co-organizers, Franklin Bost and Faisal Mirza.

So with that, I'm going to turn this over to Franklin to describe a little bit about the group that has brought this workshop to you.

Thank you.

MR. BOST: Thank you, Barbara. Good morning and welcome. I'm Franklin Bost, a Professor in the Biomedical Engineering Department at Georgia Tech. I'm also Education Director for the Translational Research Institute for Biomedical Engineering and Science.

And just a little information about TRIBES. We're about

bringing people together. We're not about doing specific research ourselves, but we are more of, except for some of our staff, a virtual organization where we put teams together of people to actually try to develop products and then commercialize the products.

So our partners at Georgia Tech, of course, and a couple of entities associated with Georgia Tech: the Advanced Technology Development Center; Georgia Tech Research Institute, which is about a $250 million contract institute; the Global Center for Medical Innovation is actually a prototyping center that operates under GMP conditions, so anyone can access their services. Children's Healthcare of Atlanta is the largest healthcare organization for children in the United States, and they also take care of all of the Medicaid pediatric patients in the state of Georgia. Emory University (we are a joint program, biomedical engineering at Georgia Tech and the School of Medicine at Emory); and St. Joseph's Translational Research Institute is actually a pre-clinical animal study facility. It's a brand new facility located on the Georgia Tech Campus.

The purpose of TRIBES is to bring these things together. And what we do as clinicians and surgeons from Children's and from Emory and from other hospitals and specialties, bring together clinical needs and get information and put them together with our biotech researchers, and then utilizing GTRI, utilizing the faculty, utilizing our students, because what we're trying to do is train the students on how engineers talk with physicians and

surgeons and what actually the development process is. They go all the way through developing proof-of-concept models. In the last five years, we've had 250 projects work through this system in the Biomedical Engineering Department.

Then for preclinical evaluation, the Global Center for Medical Innovation, as I said, provides prototyping services, clean room prototyping services, and St. Joseph's Translational Research Institute for small and large animal studies.

And, of course, the essential component is always be in communication with business, with the FDA and with the regulatory processes and requirement that we must comply with through this developmental process.

So a lot of people speak of translational research: discovery in a lab, then "bench to bedside" translation. That is a very broad space. It takes lots of steps, lots of people, lots of expertise to actually move something from the bench into clinical practice. And that's where we see TRIBES fitting to assist that process.

One of the entities that has come out of TRIBES is the Atlanta Pediatric Device Consortium. It is mostly funded by the FDA as one of the three consortiums they funded 2 years ago. There are three main projects that were funded, but we also have seed grants, and we have funded 24 projects with seed grants. And those are from independent entrepreneurs,

companies, researchers, grad students. So we've brought a board together of clinicians, surgeons, and academicians here. And Georgia Tech, Children's, Emory, and St. Joseph's again are essential components of that.

TRIBES also conducts outreach. We have a Biomedical Innovation and Development Conference that comes up in February. This is the third time we have run this meeting. It is actually to teach clinicians and surgeons, nurses kind of what do you do on the front-end of the development process. If you have an idea, what do you do with it; what do you do with it before you talk to the lawyer; what do you do with it before you really advertise it; what do you do -- advertising it meaning talking with the company or somebody else. It gives them the basis of some fundamentals of things that they should do before you talk to the lawyer, and sometimes before you actually talk with your intellectual property organization, your tech transfer office or something similar so that you will be better educated on the process and can help move the process forward. That's a day-and-a-half conference just like this conference.

Of course, one of our main outreach events this year is this MCID conference here at the FDA. So we appreciate you coming and look forward to a great two days.

Thank you.

(Applause.)

DR. MIRZA: Thank you, Barbara and Franklin. Thank you very

much.

I do want to appreciate the Georgia Tech for being our co-sponsor on this project, and I do want to thank a number of people. I didn't actually put slides up, but particularly, out of the Critical Path Research Project, it went from a vision and idea into a generous funded project here at the FDA that I was fortunate enough to receive funding for as a medical device fellow.

I want to thank Heather Stone and Kevin Troyer, who are two ORISE fellows that were hired out of this research grant. You'll see Kevin stand up if people go over the time limit, and Heather will be passing out webcast questions as they come up. And there are a number of collaborators on this project, both internal and external.

And most importantly are the two senior champions on the project, Mark Melkerson, who is the Division Director of the new Division of Orthopedic Devices, DOD, which used to be the Division of Surgical and Restorative Devices, and I'll introduce our Director of Division of Epidemiology right afterwards.

Mark?

MR. MELKERSON: Good morning. In terms of the FDA perspective, I want to start off again with, again, a special thanks. We typically with these workshops have to work without an outside partner, but that also means without coffee, without other support. So I do thank Barbara

Boyan and the Georgia Tech TRIBES group for stepping up, the Academy for stepping up and helping out with the speakers, and the Orthopedic Device forum where this idea was first launched; also the FDA staff both from Office of Surveillance and Biometrics and the Office of Device Evaluation and, again, OSMA, the manufacturers.

In terms of FDA's mission, this is one of those missions that we are to protect and promote the public health. Well, we do real well on the protect side, but we tend to act as a -- at least described as an obstacle. So this is one of our attempts to try to help promote that activity. And the other part of that, we're supposed to be facilitating innovation. I've been with the orthopedic group for 25 years as a reviewer, branch chief, deputy, and now the division director. But innovation has been a task that we've been trying to work with the Orthopedic Device forum and the Academy for quite a while. And how do we get around this standard way of looking at clinical studies? Are there other ways to look at getting new products to market and more efficient pathways?

The vision for our organization is to have -- the U.S. have access to high-quality, safe, and effective medical devices and the importance in being first in the world. I'm not quite sure first in the world is one of those that the staff quickly embraces, but first in the world means that we're getting new products to the clinicians, to the public in a safe and effective manner.

In our strategic priorities in the premarket world, we try to improve public health and foster trust among employees and our constituents. Well, this workshop is one way to bring those two together. And there's a couple of guidance documents out here. The other thing is to enhance transparency. What are we thinking about when we're trying to look at different ways of looking at clinical studies and premarket review?

Taking that to the next step, there's a draft guidance document out, which is in the process of being finalized. And it's to look at design considerations for pivotal trials. You'll find coming up there'll be the new strategic priorities for 2013; look for an emphasis on clinical studies.

But in this case, we're looking at what can we do in terms of effect size, sample size, power, patient success criteria. We've been looking at survivorship. While waiting for registries and long-term catches, the long-term issues, how do we get something to the market in a timely fashion? To do that, we're also looking at the benefit/risk, which is another guidance that has just gone final.

So in the premarket arena, we're looking for subject-reported outcome instruments. In other words, are there alternatives to just revision/reoperation. And so these are some of the things that we're looking at. And there's another guidance document out -- again, these are just concepts. Now the question is how do you put the concept into practice. And that's actually where part of this workshop focuses.

In terms of the outcome research, we're just kind of acknowledging some of the other activities that FDA has been involved in. But in this case, this is an activity that we wanted to strictly limit to the orthopedic world.

Some of the other groups that are looking at data -- and I'm also involved with our Office of Surveillance, ICOR, the FORCE, which is the Function Outcomes Research Comparative Effectiveness in Total Joint Replacement; they're doing something similar. But the question that we're looking at is what do we need for the study designs up front versus waiting for registries.

Our benefit/risk guidance has just gone final, and it's a way to try to ascertain different ways to assess benefit/risk for getting products to market. Typically, right now, that's done through our premarket approval program and our de novo, which is an automatic reevaluation of Class III. That's why we call it de novo. It's a mouthful. But we're looking at factors to consider in those determinations. One of them can be minimal clinically important differences as an alternative way to study products.

The guidance, again, for benefit/risk goes back to better predictability, consistency, and transparency. Again, these types of workshops offer that alternative. And this is a quote from our Center Director.

In terms of benefit/risk, we're also looking at the type of

benefits, the magnitude, probability, and duration. Those seem to go right hand-in-hand with some of the minimal clinical important differences. And I will get that right one of these days.

And with that, I'm going to turn it over to Faisal.

DR. MIRZA: Thanks, Mark. And, you know, thanks for highlighting the key points of why we're here. And I would like next to ask our Division Director of Epidemiology, our other senior champion on this project, whose team has been very integral to the research project, Danica Marinac-Dabic.

DR. MARINAC-DABIC: Good morning. And on behalf of the Division of Epidemiology, I would like to thank you all for coming to this very important and timely conference.

As you know, the field of epidemiology clearly is very involved in these types of research, and we would like to continue to contribute to the mission of our center and to collaborative work with all of you to advance the field and help both our premarket and postmarket missions.

Some of the driving forces that we are facing here I'm sure are the same driving forces that each of you in your areas face as well. Our patients have needs throughout the device lifecycle, from the time when the product is approved and then beyond that when it enters the clinical practice. We at the Center would like to be able at any point during this device lifecycle to be able to provide timely and valuable information to the patients and to

the clinicians, the information that is based on the best available evidence.

Devices that we regulate evolve very rapidly. And methodologically and infrastructure-wise, we would like to be able to continue to be ready to meet this challenge. This is why this conference is so important, so we can tackle some of these important challenges, talk about the needs and talk about the gaps that all of us together can work and improve the field.

In addition to the challenges that Mark Melkerson talked about, there are some specific challenges that the postmarket field faces. Particularly, these are the gaps in an integrated system in the United States, system of surveillance for medical devices. As you know, we currently have fragmented pieces; we talk about passive surveillance system; we talk about, you know, epidemiologic studies; we talk about different apparatuses we have. But what is really -- what we lack is an integrated system that will be able to serve all stakeholders, including FDA, clinical community, patients, academia, industry, and others.

The learning healthcare system as a concept I'm sure you're familiar with. Again, it has challenges and also offers a lot of opportunities to all of us. We live in the 21st century. There is a demand for a globalization of information; lots of new ideas, lots of new knowledge comes from different countries. So what we are trying to do here, and especially in the Division of Epidemiology, to try to put all these dots together and try to bring to our

Center the best available information about devices that not only comes from the U.S. studies and U.S. sites, but also from the international registries, from the international studies that are being conducted. As you know, many devices reach the market much sooner in other countries than in the United States. So we would like to be able to put all this evidence together in a synthesized fashion so that we can make, during our everyday practice, the best information based on the best available data.

By doing all of this, we will be able to contribute and lead the development of learning regulatory science, the system that it's going to learn from all these experiences, and to certainly contribute to the mission that Mark outlined in his presentation.

In addition to advancing the regulatory science, all these concepts are really helping us to develop the learning public health practice, which is again one of the missions that our Center has.

In the postmarket phase, it's very important to again keep in mind that devices are very much different than drugs. And all these differences that are outlined on theses slides contribute to some of the challenges that we have in the postmarket phase on doing the studies, doing the surveillance.

When you talk about devices, they are very much heterogeneous. Even in the field of orthopedics, you have a number of devices that have many, many different models and types. And if you can,

imagine how many of these different device types exist when you talk about the entire landscape of medical technology.

They are very often comprised of very complex components. And the way how they are developed, they go through those iterative changes, so it's not like with a drug when you develop the particular medication, and that's pretty much what it is throughout the entire lifecycle. With devices, you go through different changes. And one of the challenges of the effective postmarket surveillance system is to be able to assess those. You know, some of the discussion about PROs that we are going to have throughout the day will also focus about how we capture the assessment of these iterative changes of various technologies.

Design error is something again that's unique to devices, and our pharmaco-epi colleagues do not ever need to think about that. There are human factors in application of medical devices whether those be, you know, physicians implanting the products or the patients using the product. There is a huge learning curve for some of the devices, especially with some complicated technologies. And when we talk about surveillance, we need to think about these methodologies, how one assesses these.

Currently, we do not have unique device identification yet, but we do have a proposed rule out, and sometime next year, I think by the end of May, we will be putting together the final rule, which will give us the opportunity to start doing really device-specific research using various data

sources.

Some of the challenges in the postmarket field that are specific to devices would be the device identifiers generally are not captured or not available in health-related electronic records. So we currently are not able to systematically utilize the electronic health records for the evaluation of performance and clinical outcomes of medical devices in the postmarket setting.

Health outcomes of interest also in the field of devices, including orthopedics, lack harmonization and standardization. Device outcomes of interest are poorly captured in the records. There is a lack of specificity and standardization. The case ascertainment is more problematic without linkage of data sources. Even if the registries capture a lot of short-term device exposure and short-term outcomes, the lack of effective linkage between different data sources does not really ensure that we at the FDA have the long-term profile and knowledge about the performance of medical devices, including the orthopedic devices.

Methods for evaluation are underdeveloped for continuous surveillance, and device complexity and heterogeneity present daunting challenges to use of uniform approaches. For example, you know, the way how we approach implantable devices clearly is going to be different from the way of how we approach aesthetic devices or some diagnostic devices. So all these methodological approaches that we are working on have to be tailored

towards specific needs of the population of users for medical technology.

I wanted also to give you a context in which you're going to be discussing today the PROs and what type of tools we currently have at CDRH in assessing the postmarket performance of medical devices, orthopedics included. So we currently have in our tracking database 358 post-approval studies that were issued post-2004. And when you look into those -- and as you know, those studies are -- or you may not know, but we do have the authority to issue what we call condition of approval or post-approval studies at the time of the approval of Class III, or PMA, devices.

So as you can see, we utilize different data sources as we administer those studies. In the past, we would ask the companies to do and to sponsor their own post-approval studies that are more traditional to address a particular question that we have. But during the last several years, we are trying to explore more innovative ways of how these postmarket questions can be addressed.

And what I mean by that is that in addition to those traditional studies, where each company would do their own study and close the study after the study question was addressed, we are trying to utilize the availability of various data sources, including registries, including other data sources, including international data, also systematic evidence appraisal from the literature, and trying to address that question in a more meaningful way and in a way that the knowledge will stay for us to be able to use that knowledge

and not to really dismiss it after the completion of a particular study.

So as you can see -- Mark mentioned the registries -- we currently have 86 registries that serve as a vehicle or a venue for meeting the postmarket surveillance.

So going to the -- more specifically to orthopedic devices, we do have mandated post-approval studies for some types of orthopedic implants. For example, for hips and knees, for PMA devices, we have ongoing, you know, post-approval studies. And some of the common outcomes are outlined on this slide. I'm sure we are very familiar with the types of the outcomes we are looking at in the field of orthopedics and more commonly, you know, survivorship. Also, we are very interested in looking into adverse events. All these studies go mostly to 10 years. That depends, actually, on a device, but it goes as long as 10 years, which, again, contributes to some of the challenges because we often lose a number of patients throughout this 10 years period.

Now I would like to tie this with where are the next steps with regard to where our Center is going to revamp some of the postmarket strategies that we used in the past. This is a really very timely meeting in the sense that it follows a prominent four-day public meeting that we had in September where we launched the comprehensive CDRH postmarket strategy called Strengthening the National Medical Device Postmarket Surveillance System. Here is the actual link of the white paper that is posted also on the

web. And I encourage you to read that because it talks about -- and particularly, you know, you would be interested to look at the methodology section, and really what we are trying to do to help development of the proper methods for evaluation of medical devices.

So tied with this meeting, we also had the Medical Device Epidemiology Network Conference that followed. And also we've devoted two days of discussing the registries for medical devices. And not only device registries, but also disease registries that contain device data. Just to make sure that we understand, we are interested in all device data in the postmarket setting.

So these are some of the four pillars, or actually, the four pillars that we proposed at that meeting that will help us shape and modernize the system in the United States. I'm not going to go into details in all four, but I would like to highlight two here. Pillar No. 2 is to promote the development of national and international device registries for selected products. And I hope we are going to have some time today to discuss the value of the registries and possible application of the registries as a platform for assessing some of the outcomes that we are going to be talking about today.

In addition to that, Pillar No. 4 talks about develop and use of new methods for evidence generation, synthesis, and appraisal. And, certainly, with the patient-centeredness and the direction where the nation is going with regard to implementing the division of really having research that

is patient-centered, having the regulatory science that is patient-centered, this is also a very important meeting from the perspective that PROs and MCID is going to have its own section within this pillar, if you will, and this is a really timely meeting to talk about this.

Now, I would like to just give an example of something that is already ongoing here at CDRH. That's International Consortium of Orthopedic Registry initiative that we launched in 2011, actually, in this very room. It was an inaugural meeting in May of 2011 where we invited all international registries of orthopedic devices. And 29 of them came from 14 different nations. It was a really great approach that we put together in starting with the International Consortium of Orthopedic Registries with the mission to advance the research and improve evidence for the safety and effectiveness of orthopedic medical devices and procedures worldwide.

It is important to say that those patients, those registries capture the data on 3.5 million procedures. So here are the registries that attended the meeting, and they're now part of the consortium that is being led by the Cornell University and Kaiser, who could jointly receive the FDA contract to set up the International Consortium, that will serve as a platform for standardizing the way how orthopedic device data is captured in the registries. Again, many of these registries do not capture PRO data, but again, the importance of this meeting is to outline the needs, and we'll try everything possible that we can to make sure that those registries are

augmented to actually serve the purpose of multiple stakeholders.

In December of -- or I think it was December last year, we put together the special supplement at the general bone and joint surgery, with 14 publications showcasing the presentations given at the meeting. Some of those are captured on these slides.

We use the distributed data network. We are not in the business of centralizing and having central repository of all the data here in the FDA. We would like registries to function separately -- function independently, but also to be able to harmonize the way of how data are captured so we can do some analysis and pulling of the data as needed. And, again, this is a, again, concept of the distributed network.

And in this slide -- and again, this is -- I have probably two more slides -- I would like also to bring you up to speed with another very exciting initiative that potentially can be a vehicle of doing additional research in the area of MCID and PROs. We launched in 2010 Medical Device Epidemiology Initiative that has a mission to develop national and international infrastructure and innovative methodological approaches for robust studies and surveillance to improve medical device safety and effectiveness, understanding throughout the device lifecycle through public/private partnership with academia and other stakeholders.

And, again, the key words here are national and international and infrastructure, certainly, because I talked about a gap in the integrated

infrastructure.

The key word is also innovative methodological approaches. We talked about those gaps not only in studies, but also in surveillance efforts, active surveillance device sentinel and such.

And also what's important here, if you're talking about the evidence appraisal throughout the entire product lifecycle, not only in postmarket surveillance, which again tie to some of the mission and angle that Mark described in his presentation.

And, finally, I think what's also key is the concept of public/private partnership. This is not going to be only done by the FDA nor funded by the FDA, but through the public/private partnership, all the stakeholders will be able to contribute not only funding, but the ideas, the agendas, the gaps, and the needs. And through our joined forces, we will be able to actually tackle some of these very, very challenging issues.

We currently established a methodology center for MDEpi at Harvard under the leadership of Professor Charles Norman. We also established the MDEpi Science and Infrastructure in Cornell.
Dr. Art Sedrakyan, who was supposed to be speaking today, is the lead on this project. And, again, ICOR is already very much invested in collaborating with us in this particular field.

Some of the projects that we currently have -- again, this is not meant to actually discuss every single project, but just to give you a sense of

what type of research is being done, the evidence synthesis for CRT devices, for ventilators; we are also working with Brookings to develop the roadmap of adoption and implementation of unique device identification. We also developed clinically significant attributes for coronary stents in our collaboration with Sisters of Mercy Hospital System. We also are developing clinically significant attributes for orthopedic devices, and of course, implement UDIs in ICOR, International Consortium of Orthopedic Registries.

And we have a number of projects -- I'm not going to go into detail, but the important I think here is that we are really advancing the way how the regulatory science is done in CDRH and relying more and more on collaboration with our partners.

Some of the questions that we'll be asking you today is can pre- and post-market needs be addressed via patient-reported outcomes and can pre- and post-market needs be addressed via existing PRO instruments.

I'll certainly rely on the presentations that you'll be giving later in terms of facilitating the discussions, but to just kick off some of the thinking this early in the morning, we would like also to present our current status of the project and what we've really done in terms of our pilot that focuses on hips and knees, with the leadership of Dr. Faisal Mirza and also Dr. Art Sedrakyan from Cornell. And certainly with a lot of work of both Heather and Kevin, we went through some of the steps that we -- in terms of starting the systematic evidence appraisal in running the search on psychometric

properties for most common instruments, identify the articles for each instrument, identify what psychometric properties have been evaluated. And, again, with specific things in mind, with content validity and floor/ceiling effect, test reliability, internal consistency/reliability, construct validity, comparative instruments, comparative instruments with non-PROs and PROs, responsiveness measure, and suggested MCID value, you'll hear about these things today as we move on, and we would like to get your input in some of the methods.

And step two, in terms of data extraction and summary, we focused on demographics, psychometrics, including MCID-related items, the responsive measure, suggested MCID value and such, and provide the data summary.

And, finally, the step three, we're proposing to have the expert psychometricians' and clinicians' input on evidence summary and consensus reach-based summary.

So at this point, I'd like to stop and to thank you again for -- by coming here, showing your support for the vision that we have, again, to work together to be able to advance the field of patient-reported outcomes, MCID value in both premarket and postmarket. And I wish all of us a very productive day.

Thank you.

(Applause.)

DR. MIRZA: -- he actually was -- I was a student. I took two classes back when I was doing my residency at McMaster University, and it was an honor then, and it's an honor now to introduce him. He is Dr. Gordon Guyatt, a Distinguished Professor at McMaster University. He is the Co-Chair of the Clarity Research Team at McMaster University in which they instruct and teach evidence-based clinical practice. He has received numerous awards over the years in teaching and research, and just his brief CV is well over 200 pages long. And so I do not have 200 pages. I summarized it into a short paragraph.

He sits on many boards and committees. To name just a few, he's the editorial board -- on the editorial board of *CHEST* and policy advisor for the *Journal of Clinical Epidemiology*. He's on a number of advisory boards, the External Advisory Board of Canadian Task Force on Preventative Healthcare, and he was recently the SPRINT trial chair as well as currently on the SCOPE Steering Committee Trial for Colorectal Cancer. He is also a thesis supervisor and mentor to many students and faculty.

He has a number of areas of interest in research, predominantly in health research methodology with measuring quality of life in patients with chronic disease, effectiveness of therapy, health technology assessment, systematic overview methodology, evidence-based healthcare, guidelines development.

He teaches in the undergraduate medical program at McMaster

as well as evidence-based medicine for house staff and faculty. He has a master's and Ph.D. program in health research methodology at the Faculty of Health Sciences, where he teaches.

It's impressive how much research funding he's been awarded over the years. His total award is well over 42 million since 1983 across 117 grants. He has eight books, 50 book contributions, and numerous publications of which over 800 are peer-reviewed, and certainly almost 1,000 publications.

He has numerous presentations that he's given, but I believe this may be the very first webcast to the FDA. And also one of my friends and colleagues and now mentor, Dr. Mohd Bhandari, who is also a collaborator on this project, he has a comment about Dr. Gordon Guyatt that I wanted to read:

"Gordon Guyatt's influence on paradigm shift from eminence-based to evidence-based orthopedics has been unparalleled. He has and continues to have immeasurable impacts on the care of patients worldwide through his tireless efforts to put evidence at the forefront of surgical practice."

And certainly that rings very true. The concept of this workshop, the very first paper I came across historically on MID, MCID, was by Dr. Guyatt and a collaborator, Dr. Jaeschke. And so once we have Dr. Guyatt on, it certainly is a privilege and honor to have him present here. And the

importance of this workshop is -- it is a privilege that he was able to speak at least by webcast. He wasn't able to come in person because of his hectic schedule, and so I appreciate it.

So once we have him logged on -- I know there were some issues with webcast login, and if you have any colleagues that have e-mailed, certainly, they can simply click on as guests on the webcast. So we'll get started with Dr. Guyatt's lecture shortly.

(Pause.)

DR. MIRZA: Dr. Guyatt, are you able to hear us?

DR. GUYATT: -- so proceed?

UNIDENTIFIED SPEAKER: Yup. Are -- okay, yup, so you can proceed.

DR. GUYATT: Okay. And I speak into the telephone, is that right?

UNIDENTIFIED SPEAKER: Yes, that's correct.

DR. GUYATT: All right. Here we go, then. So I'm going to present a history and an overview of the concepts of minimal -- what we used to call minimal clinically important difference and at least what our group now calls the minimally important difference, and I'll talk about that transition in the course of the presentation.

So I'm going to briefly refer to what we think of as patient-important outcomes, the problem of interpretability when using patient-

important outcomes, strategies for making results interpretable, and I will focus largely on this concept of minimally important differences, approaches to minimally clinically important difference, or MID specification, and something called a responder analysis.

So what do we mean by patient-important outcomes? Clinicians are very familiar with looking at either survival or major morbid events, and many clinical trials focus on such events. You also have clinician-reported events and those reported by caregivers, much less often used, and what we refer to as patient-reported outcomes.

Historically, we tended to focus on health-related quality of life, but the notion of patient-important outcomes has got broader, including symptoms, global impressions, and the other domains that you see here. In this presentation, I'm going to focus on what maybe perhaps is the most important -- remains the most important of the patient-important outcomes, health-related quality of life and the issue of interpretability in health-related quality of life.

So when you have patient-important outcomes, particularly health-related quality of life, there's going to be the issue of how to interpret the results. So if you tell clinicians that for a treatment group the patients were improved by an average 5 points and there was no change in the control, how does the clinician interpret that? Is this a trivial change that they should be ignoring, a large change that mandates administration of an

intervention to the entire population, or somewhere in between?

So, for example, reports of results that might be difficult to interpret, this is an example of Alefacept on quality of life in a large sample of patients with psoriasis. And what the investigators tell us is that the drug significantly reduced, which means an improvement, mean dermatology quality of life scales compared with placebo, 4.4 versus 1.8 at 2 weeks after the last dose and 3.4 versus 1.4 at 12 weeks.

Well, in relative terms, that seems quite big, but we really would be left with questions as to whether this is trivial, small but important, or large. And this is true -- this problem is true with virtually all patient-important outcomes, all health-related quality of life measures.

So when we started working in this area many years ago, there had already been work that looked at different approaches to this problem. And one were distribution-based approaches, which rely on the magnitude of the effect in terms of some measure of variability, such as standard deviation units. So I'm presenting the results in standard deviation units, and work by Jacob Cohen had already laid the groundwork for that.

And that is one approach which I'm not going to talk about. The approach that I'm going to talk about is an alternative, which are called anchor-based measures. And what you do is take the instrument that you are trying to make interpretable and relate it to some independent measure. If this is to be successful, this independent measure has to have two

characteristics. One is that it is in itself interpretable, and secondly, that it has a substantial correlation with the new measure that you're trying to make interpretable. If either of those conditions are not met, that anchor-based approach is not going to work.

Well, what might these anchors be? They might be single items such as a diagnosis, depressed or not depressed, particular symptoms; they might be a functional classification system like New York Heart Association's functional class or the ECOG system used in cancer. That's only going to work, however, if the clinicians are extremely familiar and are able to interpret the results from these functional classification systems.

Global rating differences between individuals has been used; one individual classifies or reports themselves as minimally disabled, another as moderately, another as severely disabled. And you could look at the difference in scores in your target instruments on individuals who report themselves as minimally, moderately, or severely functionally impaired; or the approach that we took in our own research that has been widely used and which I'm going to focus on this presentation, which is global rating of change within individuals. And I'll talk about that in the next few slides.

So the term MCID, or minimally clinically important difference, was first used in a paper that we wrote in the *Journal of Chronic Disease*, which later became the *Journal of Clinical Epidemiology* way back in 1987. In that paper, however, we didn't define the concept although we referred -- we

used the label. And we first defined it in a paper in 1989 in *Controlled Clinical Trials*, where we said the minimally clinically important difference was the smallest change that patients would consider important. This is worth distinguishing from the minimally detectable difference. So this implies that there are some changes that patients will notice but that they would not consider important. And that would be the minimally detectable difference. We're talking about changes that people, when they evaluated them, when they considered them, would say these are important to me.

So in the paper in which we first defined the question, we applied the approach. And we used it with two different questionnaires, which were very, very similar, a chronic respiratory questionnaire for patients with chronic constructive pulmonary disease and a chronic heart failure questionnaire for people with heart failure. And the reason they're so similar is because although the underlying physiology is different, respiratory versus cardiac problems, the experiences and problems that the patients with chronic respiratory and chronic heart failure experience are virtually identical. The questionnaire has 20 items that look at shortness of breath, or dyspnea, fatigue and emotional function.

At the time we addressed the issue of interpretability for the first time, we had already done what was then the traditional thing is -- well, the one traditional thing about evaluating what has become patient-reported outcomes, which is look at the validity in measuring what it really is supposed

to measure.

And also at that time, it was relatively recent. Instead of looking at reliability, which tends to be very important when you're trying to say is person A better or worse off than person B, we also had established responsiveness, which in the clinical trial situation, we're very often interested in changes, are people better or worse, and responsiveness has to do with the ability to pick up small but important changes. So already at the time, we looked at interpretability. We had pretty convincing or compelling evidence about the validity and responsiveness of these questionnaires.

How do these questionnaires work? For instance, if you are asking somebody about their shortness of breath, the response options go from 1, extremely short of breath, to 7, not at all short of breath. Higher numbers, then, are better. And you ask similar questions about fatigue, how fatigued are you, or about emotional function, all on 7-point scales of this sort, where higher numbers are better and lower numbers are worse.

So in considering interpretability way back at that time, I was in the habit of going in the front lines with my research staff and sitting in on interviews and chatting with patients. And in looking at questionnaire results and comparing them in our interaction with the patients, we gained a clinical impression that the smallest important difference was about 0.5 per question. So in other words, on a domain with five questions, a change of 3 or more, improvement of 3 or more would be -- tend to be seen by the patients as

important; seven questions, 4 or more would be seen as important.

The studies that we used for further exploring this included one with 31 patients in a respiratory rehabilitation program, 24 patients with chronic airflow limitation, otherwise known as COPD, or chronic obstructive pulmonary disease, in a trial of bronchodilators, and 20 patients with heart failure in a trial of digoxin.

And in addition to administering the chronic respiratory questionnaire and the chronic heart failure questionnaire, we asked the participating patients: Has there been any change in your shortness of breath or your fatigue or your emotional function for those domains since the last time you saw us; worse, about the same, or better? In those who said they were worse or better, we went on to ask them a further question. So if they said they were worse, the response options then went from almost the same, hardly any worse, little worse, somewhat worse, and so on.

In our classification, we used responses as 1 to 3 to represent a small but important difference; in other words, the minimally important difference or minimally clinically important difference; 4 to 5 as moderate improvement; and 6 to 7 as large improvement or, with this particular scale, deterioration.

So what did we find? In the patients who said they were unimproved, there was very little change in either dyspnea, fatigue, or emotional function. These numbers are on the 7-point scale. So where 7 is

great and 1 is awful, these are very little change.

For the global rating of change that we considered to represent a small but important difference, the changes were in the vicinity of the 0.5 that our clinical impressions would have suggested. And for those with moderate or large changes, there were greater changes on the questionnaire.

So one of the things that you'd note here is, for this to work, there has to be a correlation between the global rating of change and the change in the instrument. And that's what you see here; global rating of no change, no change in the instrument. As the global ratings of change get larger, the change in the instrument score gets larger, with the small but important difference corresponding to a global rating of 1 to 3 being in the range of 0.5 that we predicted.

We then did this with a completely different but similarly structured questionnaire, an asthma quality of life questionnaire, and again, the small but important difference was really quite nicely distributed about the 0.5 that we predicted. We did that with several other instruments, and it kept being the same. The minimally important or minimally clinically important difference appeared to be in the vicinity of 0.5 when one framed one's response options into 1 to 7 scales. Not everybody else has replicated that, but it was very consistent in our own work.

So once one has established the minimally important or minimally clinically important difference, how should one use it? How does it

help you to make results interpretable to the clinicians and ultimately to the patients?

So we did a randomized trial of lung volume reduction surgery. Patients with severe emphysema, COPD, have overinflated lungs. And what might be very counterintuitive, but it was hypothesized physiologically that actually chopping out a portion of the lung could improve the mechanical properties of the lung, and as a result, the lung would function better even though there was less lung. We did a randomized trial of 55 patients who were followed for a year, and our key outcomes was the measurement of what we call health-related quality of life, the chronic respiratory questionnaire, with the main domains of dyspnea, fatigue, and emotional function that I've mentioned to you previously.

This is the abstract from what we eventually published, and it said lung volume reduction surgery resulted in a significant between group difference in each domain of this chronic respiratory questionnaire at 12 months and noted not only that the change was 1.9 in dyspnea, 1.5 in emotional function, and 2.0 in fatigue, but noted specifically in the abstract that 0.5 represents a small but important difference. And knowing that, the clinician could say, well, that 1.9 and 2.0 represent approximately four times the minimally important difference, and the emotional function represents three times the minimally important difference.

And in presenting our results, we presented that graphically.

So here is in blue what happened to the control group; they deteriorated. Here is what happened to the patients who had lung volume reduction surgery; they actually improved, less dyspnea, fatigue, and better emotional function. And the gray area here represents a graphical depiction of the minimally important difference of 0.5.

So here's a visual depiction that shows by the end in this particular domain, which was emotional function, the improvement, or the difference, between treatment and control by the end of the follow-up period of 12 months was approximately three times the minimally important difference. And we believe that presenting that puts you much farther ahead than if you simply told people the difference was about 1.5 in this instrument that they might have limited familiarity.

Well, I've referred already to the change in our terminology. And in 2005, we published a commentary specifically addressing that, which we entitled, "Good-Bye MCID. Hello MID. Where Did You Come From?" And what we said in this is the clinical in MCID, we believe, focuses attention on the clinical arena rather than patients' experience in their day-to-day lives. And that's why we got rid of the "clinical" from the MCID. Health-related quality of life is an important outcome because it's the patients who experience the outcome, and only they are in the position to ultimately judge it. And the clinical distracts, we believe, from that effect -- from that realization.

So in that paper, we made a slight redefinition and now define that minimally important difference is the smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and which would lead the patient or clinician to consider a change in management. So as in our initial definition, we tied it to changes in clinical actions, changes in what might be the optimal management for the patient.

So I'm now going to deal with some potential problems with the minimally important difference and problems with our particular strategy for establishing the minimally important difference, which was this global rating of change. In much of this work, I should acknowledge, are internal McMaster critic Jeff Norman, who kept throwing roadblocks, productive roadblocks because they were points well taken that we had to deal with, questioning our work.

So what Jeff suggested is that when people make this global rating of change and say they're better or worse, they're, in fact, not telling you they're better or worse at all. All they are telling you is today I feel good or today I feel not so good. That really has nothing to do with what they felt the first time you administered your questionnaire to them.

And Jeff pointed out -- Jeff is very -- his background is physics, and he functions as a methodologist and as a statistician. He told us something that I would never have known myself, which is if this global rating

of change is working as it should, the correlation of the change should be equal and opposite for the pre and post measures.

So typically what we do, we look at the global rating of change, and we relate it to the change in the questionnaire, in other words, time 2 minus time 1, and compare that to the global rating of change. Well, you could take just time 1 and time 2 and correlate it with the global rating of change, and if it is working the way it should, the correlation should be equal and opposite.

So we asked: Is this happening? Well, the first dataset that we looked at was 39 patients with asthma. We looked at their particular global ratings between visit 2 and visit 3, and we demanded a minimal correlation of 0.5 between the global rating of change and the time 3 minus time 2. You'll remember that one of my initial specifications for whether any anchor works, it has to be at least minimally correlated with the instrument that you're trying to make interpretable, and our criterion for that minimal correlation was 0.5.

So here was the overall score on the asthma quality of life instrument. And what you see is that the pre-correlation and the post-correlation are opposite. The post is a little bit higher, but one wouldn't expect it to be -- work exactly right. And the most important thing you see is that the correlation of change, the post minus the pre is substantially higher than the correlation with either the post or the pre, suggesting that the pre is

indeed contributing, and in fact, you're getting a rating of change.

Activity limitations, this should be minus 1. -- .13, or actually, minus .47 if it's working perfectly; clearly not working perfectly and clearly, the correlations invariably are higher at the post and the pre. They seem to be -- the interpretation psychometrically of that is they know their status today better than they knew their status when you asked them the first time, which is not too surprising. But the status the first time is contributing nicely in that you have this bump up in the correlations.

Here's one that works almost perfectly. The correlation of the pre and the post are almost equal and opposite, and then you have this big, big bump up in the correlation when you use the time 2 minus the time 1.

Here's another set from a rhinitis questionnaire that didn't work at all so well. So here you see that the bump up doesn't exist. The correlation with the post and the correlation with time 2 minus time 1 is almost identical, suggesting that just criticism in this particular dataset is, in fact, very legitimate. People aren't remembering what happened before, and they're really just telling you how they feel today.

But here's another from a different dataset. We did this on a whole bunch of datasets. And here's ones that -- here's one that's working perfectly; low correlation with both the pre and the post, but when you take time 2 minus time 1, a big bump up, and these are almost exactly equal and opposite.

Here's another one where you would have liked this to be negative, so it's not working perfectly. But still you have the big bump up suggesting that there is substantial validity; people are, to some extent, remembering their pre status when they make these global ratings of change.

So the bottom line of all that is sometimes these global ratings of change work nicely. Sometimes they don't work so nicely at all. And one needs to look -- one can't make the assumption that they're working well. One needs to address the issue and empirically look at the pre- and post-correlations if one is to be confident that your global ratings of change are, in fact, giving you insight into interpretability.

Twenty or more years after we did this, there are still issues that remain to be sorted out. All our work has focused -- when we looked at the time 2 minus time 1, we were looking at absolute changes. Some people suggest you should use relative changes.

And so let me give you an example. Let's say we have a 10-point scale where 10 is bad. So now high numbers mean you're worse off. So let's say a pain scale where 10 is terrible pain and 0 is no pain at all. Well, let's say two individuals, one goes from 9 to 8 and one goes from 3 to 2. If you are using an absolute, you would say both individuals have the same degree of change, and when you looked at your -- compared them for global ratings, for instance, you would say, okay, these people should both have the same global rating; perhaps a 1-point change would be a -- you would expect

it to perhaps be small and important.

However, if you think of these as relative, 9 to 8 is a change of just over 10%, whereas 3 to 2 is a change of 33%. So if it's working as a relative, you would expect the person's subjective degree of change to be less going from 9 to 8 than going from 3 to 2. This has not been studied empirically. We're actually in the process of trying to study it empirically at the moment.

There has also been suggestion -- we have assumed in our own work that improvement in the -- the minimally important difference and improvement in deterioration is the same. There is a suggestion that in at least some instances, that is not true. And some people have suggested, and it almost certainly is true, that different groups or contexts have different minimally important differences. However, the question is still -- I would still raise the question whether these differences are important enough, very often, to pay much attention to and might just get in the way of the desirable simplicity that is necessary to enhance interpretability for our clinical and patient audience.

So the final potentially problematic issue in applying the MID or MCID that I would like to address has to do with the threshold of what the MID is. So let's assume, as we've seen in many of our own health-related quality of life questionnaires, that the minimally important difference is 0.5 and you have a -- you do a trial, and the improvement -- the control group

remains the same and the treatment group improves by a mean of 0.25, half of the MID. Does this mean that no one benefits? That might be an intuitive interpretation. Well, what if the change was 0.6, a little bit more than the MID? Does that mean that everyone benefits? Well, in both cases, not necessarily.

So a mean change of 0.25 could mean that 75% have no improvement and 25% have an improvement of 1.0, twice the minimally important difference. That would get your mean change of 0.25. And if that were the case, clinicians often like the number needed to treat, how many people you have to treat to have one person have a important change, and that would mean you would have to treat four people to have one person to have an important change. And the world of number needed to treat, that would be not bad at all.

So we looked at how one might apply this realization in a multicenter, blinded, crossover, randomized trial of 140 patients with asthma. And we looked at salmeterol, salbutamol, and placebo. We administered the asthma quality of life questionnaire, structured like the CRQ and CHQ, which I spent a little more time on, but basically operating in the same way. And our estimates of the MID, as I've already told you, for these questionnaires are 0.5.

So we now look at the individuals whose period on salmeterol minus their period of salbutamol, the difference was greater than 0.5,

favoring salmeterol, and the period of -- and over here, you have differences of greater than 0.5, favoring salbutamol.

So what you'd say is that the net benefit is the proportion of people out here with greater than a 0.5 difference minus the proportion over here with a 0.5 difference. And when you do that, here is the mean difference. 0.5 is the mean difference just bang on the minimally important difference. When you do the proportion greater than 0.5 on salmeterol minus the proportion greater than .5 on salbutamol, you have 30%, which gives you an NNT of 3.

Here, for the limitations on activity, it's less than the MID. And some people might be tempted to conclude that the therapy, then, does not have an important benefit, but approximately 1 in 5 people had a change in the intervention compared to the control greater than 0.5, suggesting there's an appreciable portion of the population that have an important benefit.

One can also, then, apply this to a parallel group trial. So here is a respiratory rehabilitation trial looking at the CRQ dyspnea or shortness of breath, and the red line represents what happened in the treatment group in terms of their change over time; the yellow line, in the control group. Here is our threshold of 0.5, and you see more people in the treatment group than the control group have changes of 0.5, suggesting that proportion of individuals might have had an important benefit.

But one can also look in the other direction and say how many

people had an important deterioration? And you see that there are more people in the control group that had an important deterioration than the treatment group. So one can add, then, the differences in proportions between these two and the difference in proportion between the deterioration to get the proportion of benefit.

And when you do that here, here's dyspnea with a mean change of something just greater than the MID, the proportion better on rehabilitation, the proportion better in conventional care; the difference is about 20%, for an NNT of 5.

Interestingly, for fatigue and emotional function, where the change is a little bit less than the MID, tempting the naive interpretation that nobody benefits, but when you look -- when you use this, which has been called the responder analysis, you have actually a larger proportion benefiting from rehabilitation and a smaller NNT than you did with the dyspnea.

Now, a lot of times, the work hasn't been done to establish an anchor-based minimally important difference. What do you do then? Well, you could do the studies to find out. You could guess. And it turns out what exactly the minimally important difference is doesn't matter much, assuming certain assumptions are met. And the assumptions are as follows: The assumption is that the true scores in the intervention group are normally distributed, and the true scores in the control group are normally distributed. This, again, is work from Jeff Norman, or work led by Jeff Norman.

But, anyway, so here you have the effect size in standard deviation units. And the issue is: Where do we put this MID and how much difference it makes? So here in the responder analysis, if you set the MID at just under 1 standard deviation unit, then we are going to get a proportion that differed by -- we could do that -- all you need to know is tell somebody what the effect is in standard deviation units and give -- assume that the intervention and control group, their results are both normally distributed, and you can make, theoretically, the calculation of those who had more than the MID improvement in the intervention group and the control group.

So on this slide, what it shows is the effect in standard deviation units and the proportion who benefits when the threshold, the MID threshold is 0 standard deviation units, .2, .5, and .8. And you basically see that it makes very little difference what threshold you choose. And here are three particular datasets that happen to follow very nicely on that line.

Well, what if you actually applied this in detail to a real dataset. Here again is dyspnea in a rehabilitation setting. And the issue is if I set my MID at 0.5 or if I set it at 1.0, does it make much difference? And as it turns out, not here; is setting the MID at 0, 0.5, 1.0, and so on. And if you look at the responder analysis, between 0.5 and 1.0, it differs very, very little. Even at 0 it's not that different; 1.5, again, you could argue it's not that different.

So at least in a moderate range from 0.5 to 1.0 doesn't make too much difference. And that, perhaps, should give us -- make us a little

more comfortable when we make assumptions about the MID and don't have anchor-based establishment in terms of making a guess in this responder analysis.

And, indeed, we have suggested this correspondence. Here's the effect size in standard deviation units. And here, if the results are normally distributed, are the number needed to treat that you might expect, the proportion of benefit simply the inverse. So an NNT of 20 means that there's a 5% difference; an NNT of 4 means there's a 25% difference.

So what are the bottom lines from all of this? What are the messages about how we should make quality of life or patient-reported outcome data particularly from randomized trials, which has been our focus compelling to clinicians?

First of all, ideally, you would establish an anchor-based minimally important difference. Second, you would report the mean differences between intervention and control, and you would report them in terms of the minimally important difference, as I've showed in that trial of lung volume reduction surgery. But you would also have to note that if you report a difference less than an MID, you'll have the risk that this -- a naïve interpretation will be that the clinician would conclude that nobody has had an important benefit.

As a result, we suggest that you also choose a threshold which can be either an absolute or a change related to the MID and do what is called

this responder analysis, where you calculate the proportion of benefit. And since a lot of clinicians seem to like the measure, you can report it as the proportion or the inverse as the proportion, which is the number needed to treat.

So that is the end of what I wanted to say. And I believe there's still time for questions. If people have questions or comments, I would be happy to address them.

DR. MIRZA: Thank you, Dr. Guyatt. That was a wonderful explanation of MID.

So we'll change that, Heather. We'll change it to MID instead of MCID.

And so we have some microphones in the center aisle, and Vivek has a mike freestanding if anyone has any questions from the audience. So -- yeah, go ahead.

DR. HAYS: Hi, Gordon. It's Ron Hays.

DR. GUYATT: Oh, hi, Ron, how you doing?

DR. HAYS: Good. I wanted to go a little bit into your -- when you're actually applying the use of the MID and you have two groups, because I think I have a disagreement about looking at the difference between two groups and using the MID for that.

So a hypothetical example would be you estimate the MID as 4, and the control group you find a change of two points and in the treatment

group four points. In my way of thinking, the treatment group has on average reached the MID, so that the change for them is something that's minimally important.

And you were focused on the difference between the two groups. And that would be a difference of two points. And what I'd say is the control group has not achieved a change that reached the MID value of 4 since they've only changed by two points, but the treatment group has changed by four points, so that's a minimally important difference for the treatment group.

So I'm a little concerned about taking the estimates that are really based for one group and then saying and applying that to the difference between two groups. So I wondered if you could comment on that.

DR. GUYATT: Tell me, Ron, what you would interpret if the control group improved by 3.9 and the treatment group improved by 4.1.

DR. HAYS: Well, that would be very close, and of course you -- you know, any threshold is arbitrary. But I'm saying if you're going to use the MID at all, you're using a threshold. And, of course, if you get to 3.9 and you're saying the MID is 4, you might say, well, that's close enough if you want to have that style. But let's say, you know, instead of two, you could make it one point, you know, for the --

DR. GUYATT: Right, okay. So it seems to me, you have acknowledged in the answer that you'd probably say there's no important

difference when it was 4.1 and 3.9, and you might well say there's no important difference when it's 4.2 and 3.8. And at some point, as the difference between intervention and control became sufficiently great, you'd say, okay, I now believe there is a difference. As soon as you've said that, it seems clear that you have to take into account the difference between the intervention and control.

DR. HAYS: Well, taking into account is fine. I'm saying your estimate of the MID is not the difference between the two groups. It's within one group. And so I agree there's -- it's important --

DR. GUYATT: Well, but what you just said seems to me to contradict your acknowledgement that you're going to call the difference between 4.1 and 3.9 likely trivial and unimportant.

DR. HAYS: I think that's another consideration is what I'm saying. If a value is 3.9 versus 4.1, you might want to say that those are essentially equivalent values. I'm saying when you apply the MID, you shouldn't be applying it to the difference because it has nothing to do with the difference between the two groups. It has all to do with one group.

DR. GUYATT: Well, I guess we have a --

DR. HAYS: Yeah.

DR. GUYATT: I guess we have a fundamental disagreement that it would seem to me that it is very unlikely with 4.1 and 3.9 that you have a substantial difference in the proportion that have received an important

benefit in the intervention/control. When it's 4 and 2, it is much more likely that there is a difference in the proportion who have received an important benefit in the intervention and control. And when you have 4 and 0, it is much more -- very likely that there's an important difference in the proportion who received an important benefit.

So I guess we have -- that would be the way I would look at it. And if you are ignoring the magnitude of the differences, I think you're making a mistake.

DR. HAYS: Okay. I think we agree with that point. But then I'll bring up one other thing. If you want to know who benefits, I don't think you can use the MID for that either because if you look to see if the individuals changed significantly or not, you're going to find that using those MID cutoffs, or any arbitrary cutoff will not necessarily show people who have actually benefited, and you should actually be looking at a statistical test of whether the individuals changed significantly, and that would be a direct way of knowing if they've benefited or not.

DR. GUYATT: Well, you can't, unless you're doing -- trial with multiple comparisons between intervention and control in the individual. You don't have reliable estimates of what the true change is in the individual, right?

DR. HAYS: Well, you can estimate it based on the reliability of the group, or you can use item response theory to get a more accurate

estimate of the individual's standard error.

DR. GUYATT: Okay. So there are various approaches. So the logic of the way we do it is as follows: You first of all do an analysis in which you treat it as a continuous variable. So when you do an analysis of treatment of a continuous variable, that tells you whether chance is a likely explanation of any differences that you have seen. So, you know, if you end up with a confidence interval that overlaps no effect and a p-value of .3, you would stop there. If you have a convincing difference between intervention and control, then you look at the mean -- you've argued you just look at who crosses the threshold and who doesn't, and I've argued that you look at the difference between intervention and control.

So the next thing that we would do is look at the intervention and control and see how that relates to the MID. But once one has been established that there is a true difference, the challenge is to make that difference interpretable to the clinician and patient population. And we think that choosing that -- choosing a threshold and looking at the proportions less than or greater, as the work that I've showed you, if the results are normally distributed, probably you're pretty safe to cross a range that is near the true minimally important difference, and you get a proportion.

And that proportion is not designed to say is there a true benefit. That is established by analyzing the data as a continuous variable. That analysis is to give a sense of the magnitude of the effect to clinicians.

And clinicians are used to handling relative risk and absolute risk and absolute risk reductions, and it is that characterization -- proportion of benefit, which at the very least, the analysis that we suggest, I think, is a ballpark estimate of that proportion, and that makes things -- having established there's a true benefit, that makes things more interpretable for the clinician.

DR. MIRZA: Okay. Are there -- if there are no other questions, I actually have a question for you, Dr. Guyatt, and again, thank you for being here with us. I noticed that in the various questionnaires or instruments that you had presented, you described the MID relative to each subdomain. And oftentimes, certainly from the orthopedic devices at the FDA, when we are looking at the PRO, we often look at it as a total score and look at --request that the sponsor define the MID/MCID relative to a total score, where the assumption is that the total score is validated as a PRO and not necessarily the individual subdomains unless that has been validated in the literature.

I know the focus is on the MID that I wanted you to speak about, but if you could address the issue of how the -- you look at the MID relative to a total score or you feel that it's validated certainly for your questionnaires in the subdomains.

DR. GUYATT: Well, let's -- the more that what you are measuring in your subdomains is very closely related to one another -- and then if you're using a global rating of change, you say overall are you better, worse, or the same, and so on -- the more those domains are closely related,

the more comfortable I would be with using the overall. But it seems to me that it would be -- say your instrument has a physical function domain and an emotional function domain. It would seem to me quite plausible that an intervention might, for instance, have a substantial effect on physical function but might not change people's emotional function very much.

Let us picture in that situation if you said, overall, are you better. Well, what does the person -- how does the individual interpret that? Do they focus exclusively on physical function? Do they function exclusively in their emotional function? Do they do some kind of gestalt?

Well, the first thing is what is actually going on in people's heads when you do that? Well, then the second thing is let's say you're in the situation that I depicted. Then you get a -- and let's say one hopes -- one is hoping that the patients use the gestalt of physical and emotional function and it kind of averages out, and they say, well, here I am and here's the MID. That MID will be inaccurate for both physical and emotional function because in the physical, the function has been watered down by the emotional function, and nothing has actually happened in the emotional function. And what one might predict if actually nobody has changed -- you can't use the global rating of change; you need variability in change across spaces to use that. But assuming you have it, if you ask them specifically about physical function, you get a substantially greater MID than if you were throwing in what is, in effect, the random error for emotional function.

So the bottom line, it seems to me, is if you're using this global rating of change approach, it is highly desirable to -- if one expects differences in change in different domains, to focus on the domain when you ask your global rating of change and establish your MID for each individual domain.

DR. MIRZA: Okay. Great. Yeah, that helps. I mean, I think it allows us to sort of understand some of the different perspectives. So we'll definitely look at that.

And I think we have some questions on the web?

DR. BOYAN: Yes, we have two questions from the web. The first question is from Ted Rosenwasser. How would you approach a dataset where the results do not appear to be normally distributed?

DR. GUYATT: Well, when the results are not normally distributed -- do you still have the slides up? Do you still see the slides?

DR. BOYAN: Yes.

DR. GUYATT: Okay. Good. Let me go back. And so here are -- here is a particular dataset. I presumed that how would you -- how would you deal with the dataset that is not normally distributed. Well, this isn't quite normally distributed, although not bad. But one could I think clearly imagine -- say, one had -- the red line is skewed, that it's flat here and then bumps up here, one could look at a non-normally distributed. Well, if you accept the logic of the MID, you could still take the proportion in the treatment and control groups, who have made changes greater than the MID or deteriorated

more than the MID and calculate that very nicely with the original data.

So if you actually have patient-level data in front of you and have established an MID, no problem whether it's normally distributed or not normally distributed. The big problem arises when you don't have, as you might not in a meta-analysis, for instance, you don't have individual patient data or you don't have an anchor-based estimate of the MID, and then you start using standard deviation units and start making assumptions that way. Then you've got a really big problem. And with non-normally distributed data, things are going to fall apart.

So that tells us why it's -- even though across a range of different MIDs, it doesn't make much difference if the data is not normally distributed; if the data is normally distributed, it's a good idea to have an anchor-based MID because then you don't have to worry about that problem.

DR. BOYAN: Thank you. One additional question from the website from Moshe Vardi. Is it mandatory that the differences between the mean scores of the two intervention groups at the study end will be greater than the MID to prove significant clinical and statistical superiority? Can a study be powered around a difference in means that is less than the individual patient MID?

DR. GUYATT: Okay. I'll answer the first part of the question and then deal with the power issue. What I was trying to convey is, no, it is not necessary, and I will go back to a relevant slide. Here is a relevant slide

where you have the fatigue and the emotional function where the MID is 0.5 and the difference between intervention and control in both cases is less than the MID. And yet when you do your responder analysis, you have a substantial proportion who are benefiting from rehabilitation in that they have had improvements greater than the MID or have avoided deterioration greater than the MID, where the control group has not. And if you're a clinician who is used to looking at NNTs, those are pretty reasonable NNTs.

So the answer is clearly that it is a mistake that your difference between treatment is less than the MID and you conclude there's no important benefit. That's probably not a good thing to do.

Of course, the smaller that the difference gets, as we used before when I was talking to Professor Hays, the difference between 4.1 and 3.9, under those circumstances, you would say there's very unlikely to be an important difference. But the greater the gradient and the more it approaches the MID, the more likely that there is, in fact, an important difference even though it's a little bit less than the MID. So that's one thing.

As far as powering studies are concerned, my repeated claim over the years has been, essentially, our power analyses are all a hoax. And the reason they're a hoax is that very, very seldom are studies powered for the smallest difference that is really important. In fact, what most investigators do is that they say, all right, how many patients can I possibly recruit, what is feasible, and then they jig the numbers so that their sample

size calculation more or less corresponds to what is feasible.

And the bigger difference that you pick that you wish to detect, the smaller your required sample size. So although it's probably not the right thing to do if one were being intellectually rigorous, the MID gives you an excuse to power for the MID when, in fact, you probably, by the logic that we just went through, should be powered for appreciably less than the MID.

But since it's a game that all investigators play -- and as a result of grant committees, in my view -- they all collude with one another, because they know if they were really rigorous, all their grants would fail with the criteria, too. So as a result, everybody plays the game together. And if you are doing one of these studies as a result, even though ideally you would be powered for something less than the MID, you'll probably easily get away with putting in a sample-size calculation powered for the MID.

DR. CAMERMAN: Hi, I'm Lisa Camerman. I'm a statistician. So I really appreciate hearing all your comments. And I just had a couple of questions and one point of clarification.

So if I understand your presentation correctly, I think what you're saying is that your primary analysis is looking at the difference in means and that the responder analysis is for interpretation. Is that right?

DR. GUYATT: Absolutely correct. Absolutely, absolutely correct. First, analysis. So the logic is, is there a true difference between the groups in terms of a continuous variable; what's the likely magnitude; and

there's some interpretation benefits, clearly, when one relates that mean difference to the MID. The subsequent responder analysis is not to address the question of whether there's a true difference. It is exclusively to help clinicians interpret the magnitude of the effect.

DR. CAMERMAN: Okay. Well, thanks. That's very helpful. And also, when we're looking at this responder analysis, wouldn't it be important to look at the variability of the estimate perhaps with a confidence interval? I think when we interpret data from any clinical study, there's always a strong tendency to only focus on the point estimates. And I'm afraid sometimes that could lead us astray by, for example, if it's a small sample, that we're giving more weight to a particular point estimate than maybe we should be doing.

DR. GUYATT: Well, the problem is, you are -- it seems to me when you say that, you're contradicting the basic logic, because these studies typically are powered for continuous variable analysis. They're not powered for binary analysis. And as a result, almost invariably, if you now did the statistical analysis in the binary, you would invariably have -- conclude chance can explain the difference in results whereas the truth is that chance can't explain the difference in results. So that's why the responder analysis is just to give you an estimate of the effect and focuses on the point estimate of the difference between the two.

It seems to me, if you were going to do confidence intervals around that, you could not do it by simply taking the proportions, because as I

just said, it'll be hugely underpowered for proportions, and you will get spurious results that would make you conclude that you can't say there's any benefit at all when the continuous variable analysis tells you quite definitively that there is a benefit. If you were going to do that, you would have to -- you would then have to model it somehow to say, well, if we take a continuous variable -- let's assume that the lower boundary of the confidence interval on the difference between intervention and control was the truth and somehow model how that would play out as an NNT and the upper boundary of the confidence interval around the difference in continuous variables and how would that play out as an NNT, which in theory you could do, but we've never done it, and I'm not sure I'd recommend it.

It would seem to me a more sensible approach would be to say let's assume that the MID is 0.5 and the difference between intervention and control is your MID of 0.5, but the confidence interval is from 0.001 to 1.0 and the p-value is 0.049. Well, under those circumstances, the conclusion is, yes, you have met your nominal value of statistical significant, but it remains plausible that the effect is very small.

And then you would interpret the responder analysis as follows: Okay, our best estimate is perhaps a 30% proportion who benefit in intervention control. But when I look at the confidence interval around the continuous variable, it's clear that the effect may, in fact, be very, very small. And if that were the case, the difference in the proportion who benefit would

be very, very small and a long way from my 30% estimate.

So the bottom line is, I think you actually -- you're quite right. You have to -- do take the confidence interval into account. And when you look at the confidence interval and the continuous variable, you have to say what is my -- what is the likely effect and what is the range of plausible effect. And then when you look at the responder analysis, you interpret -- it gives you a best estimate of the proportion who responded, but your confidence -- but you look back to the confidence interval around the continuous variable in deciding how confident that you are of that proportion to benefit rather than converting everything to binary outcomes and calculating a confidence interval around that.

DR. CAMERMAN: Okay. Well, thank you very much.

DR. GUYATT: You're very welcome.

DR. MIRZA: Great. Yeah, thank you for the great questions.

And, Dr. Guyatt, thank you again for calling. I'll finish it up with one last question, and then we'll go to break. And I appreciate you taking the time to give us this talk.

The stress you've placed on the number needed to treat, I know in -- I see it, and it shows a relative benefit that is tangible when you can say that you only need three patients needed to treat for an effective outcome. And now that we're going into evaluating benefit/risk -- and for certain devices, certainly PMA and de novo, we have a new benefit/risk guidance -- as

medical officers, we often have to review files thinking about how we're going to fill in the benefit/risk worksheet for the final approval process of a device. And how important is number needed to treat from your perspective in the studies that you've done to evaluate whether a drug in your particular studies provides a benefit?

DR. GUYATT: Well, the first point is that the NNT is simply the inverse of the risk difference. So if one understands the two, they convey exactly the same information. So the issue is how to present things in ways that appeal to or intuitively grasp by your target audience. It's a communication issue.

And many of us think that as soon as you get two or three outcomes going in different directions, it boggles the mind to use number needed to treat, or what's sometimes called, over some people's objections, number needed to harm, which is probably more accurately number needed to treat -- to have someone harmed. But as soon as you get multiple outcomes, that's pretty tough to get your mind around.

And it seems clear also, even if one doesn't look at the multiple outcomes, for some people, it's exactly the same information to say 4 in 100 will benefit or 40 in 1,000 will benefit and the NNT is 25. For some people and for some audiences, one will be more appealing, and for some, the other will be more appealing. The psychometric work has -- a large body of work has suggested that natural units -- and by natural units, I mean 40 in 100 or 40

in 1,000 -- is the way people most easily grasp things. On the other hand, there's recent work that says, well, maybe percentages in certain circumstances are as good as natural units.

And when we -- in the great working group, we came down on natural units as the optimal. And we have a lot of clinician guideline panels say, come on, let us use the NNT. Well, I think you have to listen to your audiences and say, okay, we're a little nervous presenting multiple outcomes with NNTs and trying to grasp, okay, for one outcome you have an NNT for mortality of 100, and you have an NNT for reduction of myocardial infarction of 40, and you have a number needed to harm of profound fatigue of 30, and so on, putting that together seems more difficult to us. And presenting it in natural units seems more appealing to us.

Bottom line is that I think a lot more work has to be done in terms of presenting things to people in ways they can grasp and easily interpret. But the great working group's default is natural units, but listen to your clinician audiences, and if a particular clinician audience is very fond of the NNT, you can present the NNT as an alternative or to complement your natural unit risk difference presentation.

DR. MIRZA: Okay. Great. Well, thank you very much. We have run over time a little bit, but thank you again, Dr. Guyatt, for a wonderful talk.

(Applause.)

DR. GUYATT: Well, thank you very much. My great pleasure. I

would have loved to be there, but it was great to have an opportunity to address you in any case.

DR. MIRZA: Great. Thank you very much.

DR. GUYATT: Bye-bye.

DR. MIRZA: So we're going to take a short break. We are a few minutes over, actually, several minutes over. So why don't we take a 10-minute break, and let's reconvene here at just short of 10 after, okay?

(Off the record at 10:00 a.m.)

(On the record at 10:12 a.m.)

DR. MIRZA: All right. We're going to go ahead and get started as people filter in.

So the first session is on "Patient Variables and Predictors of Outcome." Can everyone hear me okay? Okay. Louder? Okay. How about now? Is that better? Okay.

So we're going to go ahead and get started here while people filter in just to stay on time. So the first session is on "Patient Variables and Predictors of Outcome." And we've had privilege of having a number of speakers join us.

And, first of all, I'd like to introduce Deborah Moore. She's going to be helping moderate this session and addressing the questions and answers that we have. And so Deborah Moore is from Carticept and is currently the VP of Regulatory and Clinical Affairs for Carticept.

And I would first like to introduce our first speaker, Dr. Jeff Sloan, who is going to speak to us from the Mayo Clinic. And he's going to speak to us on "MCID Methodology and Clinimetrics. "

DR. SLOAN: Well, good morning. Oh, I got this wrong already. Ah, there we go.

So what I'm going to talk about is actually a very nice setup by Gordon Guyatt, friend and colleague for years, and also in conjunction with Kathy Wyrwich and Ron Hays, who you're going to hear from later today, talking about basically how do we measure things that we're measuring about the way we're measuring things. And this one is being a little finicky. Ah, okay. Bear with me. I may have to play with this a little bit.

The basic message that I want to share with you, as you heard this morning already, there are different opinions, different methods, different approaches, and so on, but I want to alleviate your fears that one method may or may not lead you astray terribly and that you, by choosing a particular statistical method, you may up with an entirely wrong answer.

As you heard from the discussion after Gordon's paper, in particular, there are many approaches to determining an important difference, minimal or otherwise, and you know, there are problems with assessing things like patient-reported outcomes and quality of life, but there really are scientifically sound solutions. And I was talking with Kathy just before, well, a few minutes ago, and she said in some ways, Gordon's talk

took us back to 2002 when we started working together as a group of scientists to try to and set up the state of the science then. And many of the questions have been scientifically answered and worked on. But there's still, as there would be with any scientific endeavor, controversies, differences of opinion, and different ideas.

So what I'm going to talk about in the next 20 minutes or so is a little bit of history on clinical significance. And I'm going to talk about the half standard deviation for patient-reported outcomes, which is the method that what Gordon briefly mentioned is one of the distribution-based methods. And so it's a nice interrelation with his talk and mine in that I'm not going to talk much about anchor-based stuff, as he didn't talk much about distribution-based stuff.

And the key message I want to give here is that it's not an either/or. The key message that you hear out of such things as the FDA guidance and other words is triangulation, complementariness of methods. As any good statistical analysis does, looks at things like that old song called "Clouds," I've looked at things from all sides now.

And usually the data have a story to tell you. And no matter how good or how bad your statistician is, the choice of a t-test or a Wilcoxon -- I can say that because I'm a statistician, by the way -- a t-test or a Wilcoxon test, even if the answers are different, it doesn't mean that one's wrong. It means there's something funky going on with your data and you've learned

something.

And so while you may hear we psychometricians and methodologists almost argue and bicker perhaps about particular methodology and how things might change, the main message especially for the clinicians in the audience is that this is solid science, there is good stuff here, we know how to do this, please just tolerate our little eccentricities as we get into the nitpicky stuff in the same way that we tolerate your nitpickiness when you talk about how to deliver treatments effectively.

So I'm going to talk a little bit about the history of it and how this half standard deviation distribution-based approach works for not only PROs but survival and tumor response just to give you an indication that what we're doing in PROs actually can mirror what's going on in non-PRO methods or outcomes and then talk about how the regulatory bodies have come into this and made things both interesting in a positive way and interesting in a challenging way, as every regulatory activity should be.

So the real question is: How big is a clinically significant difference? And as Gordon said, you know, the definition of clinical significance has changed over the years. And one thing I want to talk about early is the idea of minimally clinically significant and non-ignorably clinically significant, because most of what I'm going to talk about is actually non-ignorably clinically significant. And non-ignorably means there's an effect size that is large enough that you can't really ignore. And that fits into that

definition that Gordon gave that somebody -- it's a size of a difference that would cause somebody to do something. That's basically what it boils down to. There may be more smaller differences that are also important but perhaps ignorable. And that's where we get into the nomenclature is what is minimal and minimally clinically important.

There's about -- I think it's 16 different acronyms related to the MCID -- MCPD, MCPID, MCID and so on. They all basically say the same thing: How big of an effect do we see that somebody is going to do something? And if you use that rather loose definition, you can then go to say we can make this as complicated or as simple as we want. We're scientists. We can do either way, right? Einstein said things should be simple but not simplistic. And I'm simpleminded, so I'd like to say I've got a connection to Einstein in some way. I certainly have the hair, if nothing else. And if you don't start laughing at the jokes, I'm going to go for an hour rather than a half hour, so just so you know.

(Laughter.)

DR. SLOAN: But, realistically, assessing clinical significance, we now know there is foundational scientific work to say if you don't know anything else, you can transfer things on to a 0 to 100-point scale, and a 10 point difference on that 100-point scale is not going to be clinically ignorable under virtually every statistically present distributional theory. It goes back to the 1700s. This is basic mathematics, basic statistics that we statisticians all

learn in our first course in statistics.

The real issue, though, underlying whether we can use this rule or not has to do with how comfortable we are with things. And what we did 10 years ago is we tried to get a group of people together and get ourselves comfortable with what we were saying. And the folks I mentioned already, Gordon and Kathy and Ron and 26 others, got together in Rochester, Minnesota before the snow flew and put together a symposium which produced papers on what we thought was the state of the science at the time. And it wasn't meant to be necessarily definitive. It was an attempt to try and bring consensus to the table.

And Gordon actually and Kathy were key people in the paper that talked about all the different methods that there were for exploring minimally clinically important difference. And I refer you to that paper in particular among the six papers because that paper actually was -- that's paper 1 here -- that paper is actually in the top five most cited papers in the history of the Mayo Clinic proceedings. So it took me a lot of citations to make sure they got that right -- no, that's not true at all.

But the point is that this was a really good piece of classical work that if you only want to come away with one paper to read from this conference -- I'm biased, of course, because I was involved in a small way in this paper -- but Gordon, as he did this morning, and Kathy and others did a really nice job of showing all the different methods and all the different ways

of doing things. And the nice thing is, as with most scientific basic ideas, the ideas have not changed, and so they're still relevant 10 years later.

So this is the group of six papers I wanted to refer you to.

But the first thing I wanted to talk about is: Why is it difficult to define clinical -- what's so hard about this? I mean, we should be able to do this for anything, right? Well, over the years, in working on clinical significance, I've come to understand that there really is one answer that I always get from clinicians when I ask them how important -- what is the size of a clinically significant difference for a particular treatment you're doing or a particular test or a particular biologic marker? And the answer I get from a clinical audience is always the same. It depends. It depends on this, it depends on that, it depends on 16 other factors, because that's what medical science is all about, bringing forth every bit of information into a decision-making process. And so we shouldn't be surprised that it's kind of hard to get clinical significance for what a patient says if the clinicians have difficulty in this way.

But the other thing that happens is there's two aspects; one I call circularity, and the other one has to do with a bit of paternalism that is largely gone from medicine now, but certainly is part of the issue.

The first one, circularity. If you ask people, especially clinicians, how much of a change in blood pressure is clinically important, what's the minimally important difference in blood pressure, again you'll get that answer

that it depends because, well, women have different types of blood pressure patterns than men, older people have different patterns than younger people. We just redefined what hypertension means I think for the fourth time a few years ago. And so we keep redefining it. And so the idea of knowing exactly what is clinically important in terms of something that's been around for 100 years is still evolving today. Well, maybe we need to lower our expectations of getting the perfect answer, perfection being the enemy of progress, another quote I think was attributed to Einstein.

But going back, way back when, I wanted to mention that about 100 years ago when they first came up with the idea of a blood pressure cuff, there was a controversy over whether this newfangled machine could actually measure anything useful. And the clinical trials that were put forward to assess whether the blood pressure cuff could measure anything useful or not involved the standard treatment at the time to reduce blood pressure, which was massage therapy, which is interesting because now NCCAM and OCCAM, the Centers for Complementary and Alternative Medicine, they're involved in a number of studies showing that massage therapy is actually good for a lot of things, including reducing blood pressure.

So the circularity 100 years later is what was gold standard then is now not so much used as gold standard. We need to keep that in mind because today's standard could definitely change. So the point being is we need to find something that is useful, not perfect. And so that's a key point to

what I'm talking about is that a lot of these methods, when they get into the statistical realm, and psychometricians in particular start talking about, well, your method is not perfect, well, there is no method that's perfect.

So what can we do to find something usable? Well, part of that -- as an example, we talk about paternalism and the idea of familiarity. Fifty years ago -- well, now I guess it's more like 25, but definitely 50 years ago, there was an assumption that patients could not measure their own pain; they could not reliably report their own pain because what my pain might be would be different from what your pain might be because maybe I've never broken my arm and you have and so you've got a different reference, maybe you got a higher pain threshold, and so on. But all those things are true of a lot of different measures, not just patient-reported outcomes.

But now we know after Charlie Cleeland developed the brief pain inventory after 25 years of data, showing that pain is actually a reliable vital sign. Every JCO-registered institution has to have pain assessed by every patient because we trust it now. We didn't trust it 25 years ago. And so that evolution of trust comes along as well.

We know that clinical significance is not simply statistical significance. I'll give you one example here that came from JCO in 2002, which I won't identify the author. All the p-values of a before and after comparative study, 1300 patients, all the p-values were less than .0001. And the stated conclusion in the trial was that clinically significant changes in all

domains of QOL were observed. Now, Gordon talked about power a little bit. The interesting part of that, when you've got 1300 patients, you've got amazing power to detect very, very little changes. So we know that a low p-value doesn't necessarily mean that it's clinically important.

And this is an example of this study, where it showed in the -- see if we can do it this way or this way -- yeah, this way -- the difference is -- and these are on 100-point scales -- are like 1 -- or, sorry, 2, 1, and this is 1½. These are all statistically significant, but these are not the sort of size of differences that you're going to change any practice on. And, in fact -- there we go -- when you look at midline and search for the significance of statistical significance, you can find hundreds of letters to journals since 1970 which says, well, you know, I'm disputing the findings in this paper; even though they were statistically significant, I would never change my practice based on those results because the results are not clinically meaningful.

So people know that. But just recently, for example, somebody in a conversation in a regulatory meeting talked about how, well, as long as p is less than .05, we're golden. I thought, wow, where are we? And I won't say who it was, whether regulatory or either side, but the concept of the ubiquitous p-value as being all-knowing is a problem that we still have to deal with today, which is kind of surprising.

In a similar fashion, going back to the pain analogy, I actually had to defend the brief pain inventory in a recent conference call as saying,

yes, we actually can -- the brief pain inventory actually measures pain, which I thought was interesting. And they said, well, how do you know that it measures pain? And it was just an interesting -- so we have to be aware of this innate presence in ourselves of not being sure that whether these things are measurable and tangible and concrete and have that level of trust.

Briefly, these are all the different types of methods for assessing clinical significance. I won't go through them in detail. There's two general methods, as Gordon mentioned, anchor-based and distribution-based. The anchor-based, as Gordon said, uses another variable to check whether or not changes in the variable are meaningful to the patient, where he talks about the global change score as an obvious method of that. I'm going to talk about the distribution-based method a little bit, and I'm thankful that Gordon took care of that.

So I'll flip over the one slide I had on MID for anchor-based and go directly into the foundation for the half standard deviation method for clinical significance. And where it came from was back in my nursing days in Canada in the 1980s -- man, that makes me feel old -- what nurses kept coming to my office with, as the consultant for a lot of their projects, were saying I've got this new measure that I want to measure something about patient empathy or patient well-being or -- it usually had to do with something centered on patients -- how much of a sample size do I need to figure out whether this measure is good, responsive, and that these two

treatments differ.

I said, well -- I said, all right, well, let me turn it back on you. You asked me for a sample size. What I need to know, just as Gordon said, what size of difference are we looking for? And being a simple-minded individual, rather than using the articulate verbiage that Gordon demonstrated this morning, I asked is the size of the difference, like, so big that it's an elephant and you couldn't miss it; is it so small it's like a worm on the sidewalk, you know, you'd ignore it even if you stepped on it; or is it somewhere in the middle, which I said is kind of like a duck, you know, using the analogy if it walks like a duck, quacks like a duck, talks like a duck, then if we can all agree on the size of the duck as being important.

So I put this huge scientific question down to a conversation about worms, ducks, and elephants, which, you know, is probably putting my whole scientific career in jeopardy. But it resonated with some folks. If nothing else, you know, some people laugh at the jokes, anyway. We weren't sure about this group here.

But the idea behind it boils down to whether or not we can come up with a rule that says, okay, yes, that's real. And so as Gordon indicated, there was a definition that basically said -- Elizabeth Juniper was the person in his group, I think, where he referred to the same paper that it's the smallest difference that would cause somebody to do something. Well, what we then did was turn this on its head and say we know from the '60s

and '70s that Chebyshev's theorem, which is out of statistics, says if you've got at least a certain proportion of any distribution is going to fall within that many standard deviations of the mean.

If you think about this -- for those of you who know quality control work, think about clinical -- I'm sorry -- quality control work, the 3-sigma rule. I mean, basically, the 3-sigma rule says any distribution will be covered by plus or minus 3 sigma, right? And that's basically what Cheby -- that comes from Chebychev's rule that says if you've got something covering six standard deviations, you've got the entire distribution. Well, if that's true, then, the range can be estimated by one-sixth -- or so the range is six standard deviation, so that's true, the standard deviation can be estimated by one-sixth of the range.

So the approach, very simply, is if we say the range of the tool is six standard deviations, and that's for any distribution, doesn't matter what type it is, well, then an estimate for the standard deviation could be used as 16.7% of the theoretical range. And we can translate any of these patient-reported outcomes into a 0 to 100-point range. So, therefore, just as a start, we can talk about 16.7%, or 16.7 points on a 100-point scale, as being a standard deviation.

And as Gordon alluded to, Cohen came up with effect size talking about small, medium, and large; worm, duck, elephant. Nice analogy, right? Because he talked about -- Cohen talked about how a medium effect

size, a moderate effect size is not ignorable, is important and usually significant, although not quite in those terms.

Using this method, then, we came up with the idea that we can combine these two things to talk about small, medium, and large effect sizes, and so then we could answer that question about the power analysis to say, all right, if you've got a worm, you're going to need this -- you're going to have -- you're going to need about 400 observations to get the 80% power. If you've got an elephant, wow, you don't need very many at all. You just need like -- to get 80% power, just about 25 people. If you are looking for a duck, then you need about 64 people per --

So the advantage of this method is knowing nothing more about the distribution of what's going on. This is a starting point that'll tell you that if you have 64 observations per treatment group, you'll be able to detect something that, for most statistical applications, would be non-ignorable.

When I came up with this method, we looked in the literature at a number of different people's work; Jacob Cohen's -- yeah, Gordon actually mentioned that half point on the 7-point Likert scale. That actually correlates to half a standard deviation. And you may recall from Gordon's presentation, he showed how those slides, how everything kind of zeroed in around the half point. And what we saw, we saw this as we went through a number of iterations of things, including working with Jeff Norman, his

colleague at McMaster, and Kathy Wyrwich as well, showing that both philosophically, physically, and psychologically, humans can detect an effect size of half a standard deviation; there's a couple other methods as well.

So these are the papers I was talking about. There's many more, all recognizing that irrespective of the technique, the answers are all similar. And that's, again, the message I want to leave you with, is that no matter which technique we ended up with, if we got down to something very simple as declaring a 10-point on 100-point scale as an important difference or going through an incredibly rigorous psychometric analysis involving modern methods such as item-response theory, we all end up in the same ballpark. It may differ a little, but we all end up in the same ballpark, which is comforting, because what it says is the data do not lie to us; no matter how clever we might be, the data have a message to impart.

So the effect size here, again, just to show you -- and this is more of what Gordon showed you as well -- we look at this smallest effect size of something that's important, talking about somewhere between 5 and 10%.

So the proposal that we put together a few years ago is that we can define a movement equivalent to half standard deviation as a non-ignorable required shift for clinical significance on any domain or individual item. It can be talked about in terms of the means, in terms of the standard deviation for the group or the individual.

The key piece here is what -- people will say, wait a minute, that's very simplistic; what if I know this about this situation? Then you know something more, and you should use that information to get a more refined estimate of the standard deviation and therefore the MCID. This assumes you know nothing, okay?

And that's a key point because critics of this method have through the years said, what if I know in this situation it's a rare disease and even if I just help people a little bit, you know, a 10% difference would be important? Okay. So you're telling me this is case-specific information that you know something more that even a 10% difference would be important here, and you've got science to back up why this general rule can be superseded, and that's perfectly legitimate.

And as always, presenting global solutions is always interesting because people kind of shoot at them and that's half the fun. This is, by the way, how they test police dogs. This is their final exam. If anybody bolts for the cat, they fail, so --

(Laughter.)

DR. SLOAN: So as part of that -- you're warming up. That's good. All right. We're getting into it. As part of this process, David Cella, a good friend and colleague said, wait a minute. There may be situations where maybe a duck is too big. Maybe there's a smaller effect. Maybe -- what about the size of a robin? So they talked about Cella's robin and Sloan's duck.

I thought it was really fun that Dave and I were having -- we've got a great scientific conversation here going, Dave; we're talking about ducks and robins. How long did we go to school for this? No.

(Laughter.)

DR. SLOAN: But in any case, what he talked about is maybe differences as small as 5% of the theoretical range would be important. And that's true. In some circumstances, especially pharmaceutical applications and especially, I think, in the sort of applications we're talking about for this conference, there will be situations where we know a lot more about the patient population, the devices involved, and so on, that a small difference is important.

So in defining the MCID as we go through our discussions the next couple days, we need to keep in mind for specific situations, there may very well be a raft of science that says this is how important this is to patients because this is -- or this is how important it is to clinicians treating the patients -- which they're not -- they're kind of statistical issues, but they're clinical issues.

And as Gordon said -- I feel like I'm quoting Gordon every two seconds. My apologies for that. But he's such a great guy; what can I say? He said that, you know, in terms of determining effect size, what sometimes happens in designing clinical studies, the answer is how many patients can we get or what's a realistic effect. These are not necessarily purely statistical

issues. And we can't kid ourselves that statistics will have the perfect answer for any of these things because statistics abut the real world, and so we have to be a little bit realistic.

So the duck is the guideline; robin is a potential alternative; buoyant waterfowl, you know, justification is case-specific.

But going down to where we are now is looking at the percent of patients achieving a half standard deviation improvement, this you can present as, as he said, a binary result where you talk about -- this is a study of looking at music versus no music relieving anxiety and other things in cancer patients. And we can see wonderful, good p-values. But more importantly, you'll see the percentage of differences between the two groups in terms of actually achieving a half standard deviation is radically different. And so we've met the criteria of a half standard deviation and something being clinically significant.

And graphically, you know, again, as Gordon said, you want something graphically simple so you can plot these things out and say here's the profile of all the different effects -- or all the different variables that we're looking at. These are the ones that achieve a half standard deviation difference in terms of the average. So you can look at either the percentages or the averages, and the results are going to be comparable, which is a good thing.

So I had another idea. How would this work in so-called hard

clinical endpoints? And it made me think of, all right, the idea of if this works so well for patient-reported outcomes, what about other endpoints? Well, that's a little trickier. Is it, really? So I followed this axiom, life giving you -- give you lemons, make lemonade and the idea of, no, to make lemonade, you actually need sugar so, you know, enjoy your lemon juice. Okay. So that one -- they all can't be perfect. What can I say?

So a bit of background math very briefly. Like, looking at Chebychev's rule, if we're looking at a survival endpoint, we can come up with the distribution for the survival analysis and get an estimate that the half standard deviation is half of the time -- of the mean lifetime. So for median survival, we can do the same thing, and it turns out that the half standard deviation is *t* divided by 2 log, 2 lifetimes.

And so what this means is I can define also for survival endpoints, if I'm looking at a median survival of 6 points, a clinical significant difference will be 3 months. If I'm talking about mean survival for a 6-month base and 1 year, it'll be 6 months as well. So I can apply this same method to all survival-type analyses as well. And I put the quarter and the fifth just for comparative purposes. And these are the results for the median survival comparing 6 to 4.3 months. If you see that size of difference, that's equivalent to a half standard deviation effect size.

So this is a way of now calibrating survival analysis -- there we go. So, for example, this particular study looked at patients randomized to

receive single-agent lapatinib or a combination of lapatinib plus trastuzumab. And women who got the one treatment, they had an overall survival of 9½ months compared with 14 months, okay, with the two treatments. The median hazard rate was .74. The p-value was .026. And although I don't recommend this, you know that some people are looking at that and say .026, that's just barely significant, kind of like being barely pregnant. And in statistical terms, you either are or you're not. You know, you're not kind of pregnant although somebody -- a colleague that deals in sexuality in cancer patients, well, there are syndromes, you know, where a woman can conceive -- never mind. I don't want to go into that. It's significant or not, pregnant or not.

But what's important about this, you could use this approach again in designing studies to say we can calibrate this, that the effect is one standard deviation, which we now know to be a huge effect size, right? We haven't been able to do that with survival very easily. Usually in designing survival analysis, it's a combined clinical opinion as to say how much of an advance would be important. In pancreatic cancer, boy, if you could get a couple weeks, you could get an ASCO plenary presentation, as happened a few years ago. In lung cancer or breast cancer, in particular, if you're not seeing a substantial impact on median lifetime, that's not going to be important.

So a review of the literature shows that -- something that

Ian Tannok did, he reviewed 321 randomized clinical trials. And his survival effect size ranged all over the place. So there was no consistency previously for these hard endpoints. The average effect size was .18 to .3 standard deviations depending upon tumor type. It was generally smaller than a half standard deviation, which said that what we were doing was designing studies for smaller effect size than we're expecting for patient-reported outcomes.

Similarly, you can do the same thing for tumor response, which I won't, in the interest of time, bore you with great details. Just to go right to this graph or table, I should say, where it says if you're looking at something that's got a tumor response of 50%, a half standard deviation is going to be 25%; if the tumor response is 25%, the half standard deviation is going to be 12%. So, again, this whole method can apply.

So in conclusion here, this half standard deviation method -- for a more ready interpretation of the clinical significance of survival and response studies. You can do cross-comparison studies, which is one of the most important things when we're talking about a class of studies, as we are here today. Again, it's just one method, one approach. But before you convince yourself that it's perfect, you have to know that others have come before you with other ideas.

Meanwhile, the regulatory agents got involved. And what happened, we had some committees. We did a lot of arguing over anchor-based and distribution-based. And the MID was not included in the final

guidance because the terms -- interpreted inconsistently. Responder definitions is the state-of-the-art we're looking at now. And as Kathy Wyrwich, as you'll hear her later, said the MID is somewhat dead although, as Gordon said earlier, it's dead but come back to life because it's all coming back together.

So am I allowed to take 30 seconds to wrap up?

DR. MIRZA: Sure.

DR. SLOAN: All right. Why don't we do that. So where are we or where are we going? This is -- by the way, this actually was a trick elephant that they were trying to teach to drive just so you know it.

This you probably have seen before, the FDA guidance looking at the interpretation of data, the important circle. What the motivation for the responder analysis is that we'll need to find a threshold for the change from baseline in the continuous variable and then define a patient to be a responder or not. And then the FDA reviewers will reevaluate any context of each specific trial, and so it'll simplify analysis and interpretation. And there's a lot of examples.

So I think I will actually just end there other than saying this is what the cumulative distribution function looks like to show that we're looking at the change from baseline over all possible changes and the difference between these things.

But the bottom line with all of this -- and I'll skip over the

remainder of my slides that have to do with responder analysis, which is just three slides anyway. The key with this is we now -- where we are now is that the FDA in meetings like this and others are reconvening to define clinical significance in various committees, and I think this is important work. And I'm very thankful to TRIBES and the FDA for this invitation and this opportunity to talk about this over the next day and a half.

The Mayo clinical significance group that I showed at the beginning, we're actually reforming, and we're going to try to look at the state of the science 10 years later to see if that'll help the discussion as well. We may help, we may hinder, but we're going to try. Stay tuned.

But finally, hopefully, we can work together to come up with a practical solution for all concerned. And you're all supposed to go "aw" at this because this is a romantic story.

(Laughter.)

DR. SLOAN: This is a boy mouse that is so desperate to give his girlfriend a flower that he enlists the aid of a good friend. And that's basically what we're trying to do, right? So we're the boy mouse, and the patients is the girl mouse. That's the way I'm looking at it. Thank you for your time.

(Applause.)

DR. MIRZA: Yeah, we're going to do questions at the end. And thank you, Dr. Sloan, for a great talk. Please have a seat here.

DR. SLOAN: I have to stay up --

DR. MIRZA: Yes, you do.

DR. SLOAN: So I have to behave now.

DR. MIRZA: And so our next talk is by Dr. Aileen Davis from the Toronto Western Research Institute, and she will be giving us a talk on "Patient Predictors of Outcome."

DR. DAVIS: Okay. Good. So I have no conflicts in relation to this. And so as was mentioned, I was charged with looking at patient factors as predictors of patient-reported outcomes in the context of orthopedic devices. And I'll tell you right now I was blown away of what is not in the literature.

So in looking at this, I'll talk a little bit about what patient factors, also what constructs of PROs in relation to orthopedic devices, and then they're a little bit lumped together here, as you'll see, as I go forward.

So generally, when we talk about patient factors, we think about age, sex, race, obesity, socioeconomic status. I'll tell you right now, in the context of orthopedic devices, I found about three studies that looked at race or socioeconomic status in a way that would potentially have usable data, so I'm not even going to try and talk about them.

Okay. There we go. In terms of the patient-reported outcomes, generally, pain, physical function, higher demand activities particularly for the sports clientele, so these are standardized measures that would be things that you hear names of like the UCLA measure; the Cincinnati measure; the IKDC,

but they're standardized outcomes as well; and then health-related quality of life and, as well, some satisfaction.

There's issues around whether people are using disease-specific or a conditioned-specific measure versus a generic perspective.

And then when we get into this literature, although this is about MCID, which at this point in time, we've been talking about something that's predicated on change in measures, in actual fact, there's almost no data in relation to patient-related factors and change measures. It's all based on a status usually somewhere between 1 and 2 years post-intervention.

So we actually conducted a literature review. And we were looking for articles that used those standardized PROs that we knew measurement properties, orthopedic devices, and we're talking about things that were implantable, in this case; and then patient statistical analyses where we at least had correlation coefficients or some kind of regression method. And I had to drop my criteria significantly for some cases that we included articles that had at least 30 patients.

So we did a fairly standard approach to this with our databases. We started in 2000, and we used the search terms that you can see here, including patient-reported outcomes questionnaires, treatment. And we did this separately by anatomical site for spine, hip, knee, combined foot and ankle, combined shoulder and elbow and wrist and hand.

And, again, a very typical approach. I went through all the titles

and abstracts, we pulled everything we thought was eligible; if we couldn't tell, we pulled, still, the full articles. We checked the reference lists, and then we did our abstraction to get the data.

There's got to be a special spot you hit, and I'm not finding it. So reviewed over 10,000 abstracts, reviewed -- pulled over 500 articles. And at the end of the day, across all anatomical sites, we had 126 articles. And I will tell you right now this is a narrative review. There is so little homogeneity across these articles that it's almost impossible to try and come to any kind of conclusions, although I'm going to try.

The outcomes construct, be it pain, function, whatever, is very variable even within something as common as total hip and knee joint replacement, and you'll see that probably the WOMAC usage has been the most complementary. Almost universally, we saw composite summary scores, which would combine things like pain and function, so the WOMAC subscales would be combined. And then we had some data on pain function, those activity measures. The health-related quality of life and health status pretty much is SF-36 data, although there's a bit of the NHS and some other measures. As I mentioned before, almost everything is final status measure, very little change data. And a little bit long-term data in terms of some five-year outcomes, but most is in that range of 1 to 2 years.

In terms of the patient factors, the data again is very variable. We have papers that looked at age and sex. Some looked only at age. BMI or

obesity measures were used frequently in and of by themselves. So we ended up with all these sort of univariate approaches. And, really, the primary hip and knee replacement data and some of the cartilage techniques were the only ones where we had multivariable approaches. And we had a lot of papers that talked about adjusting for age and sex or where we had matched situations but we had no data we could look at.

So I'm going to skip that slide and actually use the example. So this little gizmo doesn't like me at all. There we go. Okay. So I'm going to present you a series of slides by construct, and they're all set up in the same way. And then I'm going to do a summary slide of all the constructs.

So what you can see, this is for pain. So down the side, we've got the anatomical site, and then we've got the conditions. So the ends relate to the number of patients or number of papers for the anatomical site, not for the procedure. In the age, sex, and obesity columns, those denominators in brackets represent the number of papers that looked at the variables. And where I've got a down-going arrow, that means that the effect was worse. So, for example, females with a down-going arrow, females generally reported worse outcome.

So what you can see by looking at this table is that there's very little data around spine fusion that we can say very much of anything. Perhaps for primary total hip replacement, there's some suggestion that females have worse outcomes, but I've asterisked this because some of these

papers talk about the fact that women had more severe pain when they started, so it's really actually quite hard to interpret that final score, and they haven't looked at it as an interaction term in any way.

For the resurfacing of hips, there's some suggestion that older people are reporting worse outcomes, and older in this case generally is over 55. Maybe a little bit of suggestion, but again, small, small numbers of articles.

In the knees, where we look at cartilage procedures, so these are osteochondral allografts, AIC-type procedures, too little data to really say anything. Even in primary total knee replacements, there's not a lot of data, but it would seem that age and sex here don't seem to have a lot of effect, and it seems pretty much inconclusive in relation to obesity. In terms of the unicompartmental knees, very little data.

Function. Some of it may potentially be a little bit interpretably different, but the numbers are so small, you know, you're saying that two of three articles is telling us something versus one of three articles. So there really is not a lot of data here in terms of the spine and hip, the knee or the foot and ankle, again. One thing I will point out is notice those asterisks again, though, that seem to suggest that although females may be reporting worse outcomes, they seem to start worse off to begin with.

If we move to the composite scores -- so these are going to be measures like the Oxford knee score, combined WOMAC score, the DASH, or

disabilities of the arm and hand, that also combines pain and function. And we again see small numbers of articles. The only one that I think that may give us a hint of things is that radius fractures where they've been plated volarly, seems like there doesn't seem to be any impact of age or sex. And perhaps, as well, in the total knee replacement, primary total knees, it seems like there's no effects for age and sex for these few papers.

And finally -- or not finally -- but the activity and sport, again, small, small numbers of papers. The only one I would point to is the knee cartilage procedures where there seems to be a sense that older individuals do not have as good an outcomes in terms of their sporting activities. And in this case, the papers generally were looking at about 25 to 35 years, that once you got beyond this, you had a worse outcome.

There we go. And, finally, the health-related quality of life and health status measures. Again, similar kinds of numbers of articles that looked at this. And in relation to these -- for sure the SF-36 data -- this was the physical subscales or the PCS component score. We see a little bit where there's some suggestion, potentially, that females are doing a bit worse in terms of quality of life after total hip replacement. The knee procedures, we're seeing that sex doesn't seem to be having an effect on quality of life in six of six papers after primary total knee replacement. And it also seems that there doesn't overall seem to be an effect of obesity although -- and most of the obesity literature is looking at morbidly obese. There's a few studies that

are continuous data, but generally, it's that morbidly obese category that's been considered.

So if I try and put this together, and I must admit, I hesitated to do this, but if you sort of look irrespective of any of these procedures, and you look across pain, function, the composite scores, the activity and sports and the health-related quality of life, there's really not a whole lot we can conclude in being quite convinced that age, sex, and obesity are having an effect. I think perhaps one of the most important things here is to again keep in mind those asterisks, where in the papers where we are seeing data on females reported as being worse in a number of them, they're talking about baseline scores.

So in summary for this data, irrespective of the PRO being used or the construct being evaluated and the device, I really don't think these data support us being able to make claims about age. Sex, I similarly don't think we can make a claim, but I think that there's something interesting going forward we need to think about in terms of this baseline score female phenomenon, and I think there's other bodies of literature that would support that that needs to be looked at. Obesity, there's not really clear relationships either, although there's a little bit of data perhaps around some of those function and activity scores for the morbidly obese.

So in summary and conclusion, there really are very few data on a given device or a given construct in the way that the patient factors have

been looked at to draw a conclusion. The other thing is, these are basically postop status, and they're variable follow-up. And I would be very hesitant to say that anything we do think we might interpret out of status-based data can be similarly applied to data where we're looking at something based on a change score, such as an MCID or an MID.

So I think we've got a lot of work to do. My take on this is that we really have yet to understand the association of these patient factors with patient-reported outcome.

Thank you.

(Applause.)

DR. MIRZA: That was a great talk.

Our next speaker is Dr. Laura Tosi from the Children's National Medical Center. And I know we're running 15 minutes late, so we'll just try and keep the sessions on time, and we'll shorten our question/answer period.

Thank you.

DR. TOSI: So for those of you who are statistically challenged, I have one message: Don't give up on sex. We can figure that out, hopefully, if nothing else.

I have primarily political disclosures. I am interested, despite your statistics, on the influence of sex and gender on musculoskeletal outcomes -- and I will make this work -- because the epidemiology -- and I think that's what's different from what you just presented -- do suggest that

females suffer from injury and disease in different ways than men do, which may be part of the problem with what you're presenting. And I would argue that recognizing -- and hopefully with the help of the folks in this audience, we can start to measure better the sex-related differences that are critical to optimizing patient care.

Now, historically, in most medical terms, the analysis of sex and gender has been limited to what we call bikini issues, and people have only focused on breast cancer and reproductive health. And, in fact, pushing people to stratify their data by sex outside of these issues is a relatively new phenomenon. But when you start to do it, you end up with this incredible laundry list -- and this really -- the list I threw up here is really -- just barely scratches the surface, but you end up with this long list of disorders where we believe, or we know, the epidemiology is different, that men and women develop these disorders at different rates. And I would argue that everything else will eventually be demonstrated.

And so as you, who are much smarter about MIDs than I will ever be, start to develop the tools, I would argue that it will be very important to look at the basic biology, the incidence of disorders, how things present, how people respond to treatment, and perhaps most important politically, access to care.

So for those of you who are not wrapped up in this -- the terminology here, please remember that sex is determined in the womb and

defined by an individual's chromosomes. Gender is a very different issue. It's how you present yourself to society, and particularly, depending the society in which you live in, it may have a huge impact on whether you have access to food, education, exercise, et cetera, which may in the long run influence your musculoskeletal health.

So here's a fun overlap. We have a bunch of healthy young women running. Okay. That's the sex. But they have chosen to do a gender-based thing, namely wear high heels. Had they chosen to wear men's shoes, they would be impacting their feet and damaging their feet fairly evenly, such as the picture in primarily blue. By choosing to wear high heels, they are now having a gender-based influence on their sex-based musculoskeletal health. Kind of fun and sort of teases out a little bit some of the nuance that can be very difficult. So we have sexual dimorphism, things occurring in two distinct forms, and we have sexual dimorphism, gender dimorphism.

Going to try to speed up a little bit. And just a little memory tweak for those of you who are a number of years away from biology, remember that sexual dimorphism comes both from hormone response of genes being influenced by the hormonal milieu in which they are performing, and also, that boys and girls have genes which may or may not be located on the X and Y chromosome. And this will influence the proteins that they encode.

Now, I will make this work. I'm going to just quickly go over

three different conditions where I think that the importance of not giving up on sex are important. My particular interest in life is in bone health. And here we see tremendous sex differences in how people make their bone density, their bone geometry, how that impacts fracture risk and even post-fracture morbidity and mortality.

So we know from mouse studies that depending on sex, you will form very, very different sized bones, both in terms of overall size and cortical thickness. We know that boys will have a growth spurt later than girls and that they will close their growth plates differently and later and that this is very, very specifically influenced by estrogen. The female cells produce more estrogen locally than the male cells. The males convert testosterone differently than the females, convert testosterone to estrogen. Males have fewer estrogen receptors. And the estrogen signaling is different in male and female cells. It translates to boys grow longer.

That then influences peak bone mass. The males are forming bone longer. They make bigger bone for much more time. As they then age, this impacts when folks hit the facture threshold -- the boys cheat, too; they don't go through menopause -- so that they hit the fracture threshold after they're dead. And the well-being of bone, the risk of osteoporosis, the risk of having a fracture, ends up being very, very different in men than in women. And it translates into a woman having an almost 1 in 2 risk of having a fracture in her lifetime where in a man it's closer to 1 in 8.

Another part of the influence of sex on skeletal health, or the influence of skeletal health on mortality, is the incredible rate -- work by Werin (ph.), et al., now almost a decade old, which demonstrated that even though the men are fracturing at a much lower rate, 1 to 2 years later, a hugely higher rate of men are dead and that men develop -- and this speaks to the underlying well-being of the individual -- they develop different complications.

So as you can see from this table, men are much more likely to get pneumonia, kidney failure, or septicemia, so that 2 years after a fracture, 1 in 5 women will have died; 1 in 3 men will have died in general. So again, sex matters.

Now, a number of you are like me and you were around when Title 9, in fact, came out. And it has interestingly enough significantly changed the incidence of sports injuries in America. Girls were not playing particularly soccer back when I was in college, and now I don't know any young woman who hasn't at least tried it. But this has led to a much higher incidence of ACL injuries of the knee. We find, first of all, that these injuries are different in girls. The incidence of that injury goes up quickly at 14, peaks at 18, and then drops like a stone. In boys, they start happening later, peak at 18, that's the same, but only minimally declines through age 40. Is that because people are playing sports at different times in their lives, or is there something biological going on here?

This will work if I hit it enough times. I know it will. Many folks have tried to look at what are the different causes; why are the girls having such higher rates of injury? And one hypothesis is that estrogen is influencing collagen synthesis and degradation, and there is some sense that the tears can be correlated with where girls are in their menstrual cycle. Other work has been neurological, showing that girls land differently than the men do. And, in fact, as you look at most of the prevention strategies going on right now, it's not keeping girls out of sports depending on their menses, but it's teaching them how to land.

And Jimmy Sauderback (ph.) put this extraordinary diagram together a number of years, which again gets to the great variety of the different ways that sex and sex hormones can possibly influence whether that injury occurs; very hard to control for and analyze in your MIDs.

And I would argue from a general health standpoint for our aging population, this will be very important to do because we see that the girls who have had their ACL repair, ironically, have a higher incidence of developing osteoarthritis as they age. And there is virtually no research that has been done on what are the better ACL reconstruction techniques by sex. And we care that those girls get that.

Last but not least, osteoarthritis. Now, this is a little bit of a tough table, but the bottom line is men are more likely to get hip OA; woman, knee OA, by a lot; and hand is sort of middle of the road. OA significantly

limits our elderly population. Again, should we in the studies that you referred to be comparing men and women, because the anatomy is so incredibly different, not just the anatomy that you see across a room, but even at the level of the cartilage thickness. Women unfortunately, on average, make thinner cartilage that needs to last a lot longer. They end up reporting arthritis more, and they're much more heavily impacted by the obesity that they unfortunately have more of.

Getting to this issue of sort of societal issues, there's also gender bias in who gets total knee therapy. So everybody I'm sure in the room remembers the study in the *New England Journal* many years ago when Afro-American and white men and women went out and complained of heart disease. But a smaller, cheaper but, to me, much more fascinating study was done in Canada, where a man and a woman with identical histories and x-rays visited 71 physicians complaining of knee pain. To the embarrassment of my profession, orthopedic surgeons were 22 times more likely to recommend total knee replacement to a man with moderate OA than to a woman. And yet when they were interviewed, the surgeons could not admit gender bias.

So again, to your issue, are we comparing apples to oranges, there's a lot of feeling that women have to be much sicker and in much greater trouble before anyone will offer them treatment.

As we look at MIDs today and in this group, the issue of gender-specific implants has come up over and over. And to date, none of these

implants have been shown to definitely have a clinical effect. And yet as the owner of a mother with a male knee replacement, I would argue I think they were asking the wrong questions, because I get complaints about that knee replacement every single day. And I think it just means we need to do more work.

Finally, we see that right down to the biology of the articular chondrocytes, there are differential respondents to estrogen. And interestingly enough, there are fewer estrogen receptors in male cells. Does this make a difference? And I'm running out of time, so I'll rush along. Will this impact cell therapy and tissue engineering? I think absolutely. One of the great memories of Barbara's and my careers is doing a talk on -- or doing a program on sex difference. And Johnny Huard, who does stem cell therapy in muscle disorders, had to prepare a talk for us. And his great message was, "I'm so glad you asked me to do this talk. Up until now, I never realized that only the female stem cells would grow; those male stem cells were useless to us." So there's a lot going on there.

So as we go forward, it isn't that we need, I suspect, totally different instruments for men and women, but that as folks like yourself start to look at the results of the instruments, it may be that the women and men need to be looked at in different places, perhaps, on the scales, and we certainly need to know the sex and hormonal status of the patient in planning any therapeutic strategies. As we go forward with tissue-engineered medical

products, again, tools like you're developing here, we hope, need to be optimized for males and females and I suspect differently. And cell therapies, as Dr. Huard discovered, may need to take donor sex and patient sex into consideration.

So let me be almost on time. Thank you very much.

(Applause.)

DR. MIRZA: Great. Thank you very much. And, actually, that paper that was quoted is actually -- I'm actually going to consider a landmark paper -- is Dr. Burkhoff's (ph.) paper out of Canada that showed the gender bias. And I think that was a brilliant paper reflecting that.

We're now going to move on to Dr. Bob Campbell who is going to speak to us about the pediatric issues related to patient variables and predictors of outcome.

DR. CAMPBELL: Well, thank you. I'm a pediatric orthopedist at CHOP, but Barbara asked me to speak to you as a former PI of a 14-year FDA-device trial for a pediatric device. Don't have much in terms of conflicts. I running what's called – the Center for Thoracic Insufficiency Syndrome. It's a multi-specialty group at Children's of Philadelphia.

I looked at this conference goal, and this is tough from a pediatric viewpoint, especially for patient-reported outcomes, because a lot of my patients can't talk, so it's difficult -- but I want to show you how you can indirectly get around that.

As far as truisms about children, we pediatric orthopedists are lucky. Most kids do get well regardless of what we do for them, and intervention doesn't have to be too invasive. But there are tougher issues, and that's what I want to talk about today, where we do need devices; we need better technology. There's a child with dislocated hips from neurogenic problems, and this x ray shows a blade plate fixation. That device has been around for 60 years. It hasn't changed. It's difficult to put in. You've got to get the bone at the right angle to get the hip back in the acetabulum. It's got to not fracture out of the femur. It's got to miss the growth plates. And it really needs to come out later, but we have no resorbable devices. So there's a lot of things that need to be done here.

Skeletal maturity is a critical issue for us with devices. It defines when we switch from pediatric to adult devices when we don't have to worry about those growth plates anymore. And Laura Tosi showed you about that, but for males, skeletal maturity is a little bit older than for females. But right now, regulatory-wise, FDA defines adulthood as greater than 21 years. So we're a little concerned for those skeletally mature patients of ours that are regulatory immature, will we have devices to treat them? So I think one thing we have to work on very hard is to define that better.

Now, what I really need to discuss, what Barbara asked me is to talk about the VEPTR HDE request for approval, which I ran for quite a few years. This VEPTR device is very simple, hooks to the chest wall -- to expand

the chest and correct spinal deformity without harming growth. And it all started back in '87. I was shown this chest x-ray by Dr. Melvin Smith, a general surgeon, who has since passed away. And this is a youngster from Houston. And he was ventilator-dependent, no options, and it was felt – he was going to die soon.

We looked at this, and there were things to consider . You see a child with no chest wall; the lung is flail; scoliosis, if he lives, will make things worse. So what do you do?

So a few days later, we took him to surgery, and with some orthopedic fracture pins, we constructed a chest wall for him. And to our delight, he got off oxygen five days later and went on to grow and thrive. But it was nightmare to put in. It was not safe. It was thick steel. I had vice grips trying to bend it around the ribs next to the blood vessels, and one slip would have been catastrophic.

Well, he went on to thrive. And I promised Dr. Smith I could develop something for him that would be easy to implant and could be expanded without patient surgery. But in all honesty, I had no clue how to do this.

(Laughter.)

DR. CAMPBELL: So this starts what turned out to be a 17-year saga. And I want to tell you the lessons I learned and what the FDA learned during that time period. It started out as a custom implant. And we

developed it. It was a simple device. You could wrap it around the ribs like the old Steinmann pins, but it was expandable with a limited access out-patient surgery.

And well, after a while, the volume was too big for customs, so I called up FDA, and there were two junior reviewers, Ted Stevens and Mark Melkerson, who is involved with this, and we worked on the amendments of the protocol. And, you know, we got -- we were cut a lot of slack by the FDA. We had to invent ways to measure the radiographic outcome. There's no way to measure thoracic deformity, so we came up with what's called a symmetry ratio.

And I asked Mark years later, about the requirements for the amendment, where he wanted to know the temperature of the ventilator gases. And I asked him what was the purpose. And he said, "Well, you know, we sort of threw the kitchen sink at you." And, you know, in fairness, they needed to, because we didn't know what we were doing, so we had to look at almost everything. Now, you don't have to repeat that, for every device, but when you start out with projects like ours , that's the way you've got to go.

So we started the feasibility trial at our place in San Antonio in 1992 and then went on to a multicenter study. Why? We wanted to make sure the other guys can do it to.

We changed the device mid-trial. And that was a little bit uneasy times because it might start the regulatory clock over on us again.

But, well, the walls of the device were migrating through the bone, so we redesigned it so it had a little shoe to support the rib with a broader surface area. And then we modified it for versatility. This slide shows a very curved device to expand the chest. You need it for narrow chest syndromes. So that was important. We're glad that FDA could do that for us.

We started teaming up with the Synthes Spine company, and they modified the device to make it more surgeon-friendly and started our multicenter trial. And we looked at all these patients. It would take forever to do a separate trial for each type of anatomic deformity, but we united it on what we call thoracic insufficiency syndrome. That's the inability of the thorax to support normal respiration or lung growth. And, clinically, –we figured out this is what was killing the children long-term. The inability to support normal respiration depends on the diaphragm and also the chest wall motion – muscles of respiration. And also lung growth; if the chest doesn't grow, there is no lung growth. It usually peters out about age 8 years, but recent data has suggested that it may proceed all the way into adolescence. So that may change things in the future.

We looked at, in our multicenter trial, the indications for VEPTR, and you'll see them here, and it's a progression of the chest wall deformity both on x-ray and also clinical progression. We didn't have any hard and fast criteria. We had to look at it clinically and to address it we used these procedures to stabilize or enlarge the chest. And we felt it would allow

growth of the spine and chest, which would be helpful for lung enlargement, and also, clinically, we think it would help the patients.

Well, how do you measure that? Big question. Well, radiographs. We orthopedists are very happy with radiographs. And you look at this infant with a resected chest wall tumor treated by VEPTR, so the scoliosis has improved 20 degrees. But what actually is going on? It's hard to tell on xray, but at least we see where the devices are. CT scans show a great volume addition to the chest that xrays could not show.

We needed a quality of life assessment that really correlated with these children who were on ventilators or oxygen. We came up with what we called assisted ventilation rating. And it's an ordinal scale. And, you know, we just based it on clinical experience, and the higher the number of the scale, the worse off you are. And it's an exponential increase in quality of life; when you can drop this score. With full-time ventilator support, if you have a respirator hose come loose, that child may be dead in 4 to 6 minutes. It's just nerve-wracking being around a child with that sort of, you know, frailty. So if you could move them down to part-time ventilator dependence, if the hose comes loose, he will survive, and you can get to him in time.

We did look at what was available at the time. This is back in the late '80s, you realize. And we looked at the Infant and Toddler's Child Healthcare Questionnaire for kids under 5 and also Child Health Questionnaire for those over age 5, they turned out not to be too helpful.

And if you think about it -- that's my daughter there on the slide -- I could not function as an accurate spokesman for her on a quality of life questionnaire, and I think she would agree.

So you can look at this -- with this Child Health Questionnaire, it was analyzed later by the Columbia Children's Hospital Group and Dr. Michael Vitale in New York. And they looked at the VEPTR data, and concentrated on two things, the functional status/ psychosocial well-being, and the burden of care on parents. And he looked at 45 of our patients and pulled the FDA trial data. For the preop thoracic insufficiency patients, physical scores were lower than any child in the literature, including those with cancer. These are very sick children. And they had very high caregiver burden scores, and this is due to the child being sick, the emotional problems, learning abilities, everything. It just puts a big, big hit on these parents. Looking at the postop scores, about nearly three years afterward, there was no significant improvement in quality of life for these children and burden of care after that. Now, that disagreed with what we saw clinically. It appears to stabilize quality of life, if you want to look at this way, and maybe that's the best we can do.

There were a lot of unknowns in these situations. We didn't know about other treatments or surgeries for comorbidities. There were unknown family dynamics. You know, patients with severe disabilities -- parents seem to often get divorced in these situations. What really had an

influence was -- unclear natural history also. There was no significant difference, however, between those with complication versus those without except for parent distress, and that's probably over concerns or financial stress. So there's a lot of work that needs to be done in these areas.

And if you look at the quality of life instruments available for scoliosis alone, older adolescent patients, there's just a slew of them, which implies to me there's no gold standard. Dr. Fertali has come up with an early onset scoliosis questionnaire. EOSQ24. (Measuring quality of life in children with early onset scoliosis: development and initial validation of the early onset scoliosis questionnaire.
Corona J, Matsumoto H, Roye DP, Vitale MG.
J Pediatr Orthop. 2011 Mar;31(2):180-5. doi: 10.1097/BPO.0b013e3182093f9f.)

What about the VEPTR trial outcome measures? We also looked CT scan of the thorax. You look at this x-ray, it doesn't tell you much about thoracic insufficiency syndrome. You look during surgery, there's an obvious increase in volume. And you look at the CT scans, you can tell where the devices are, but you see the right lung that was collapsed in this patient has been expanded by the procedure. And this is CT scan work from Boston Children. You can see the increase in lung with CT scans. We also looked at echoes, capillary blood gasses, pulse oximeter; didn't turn out to be too helpful in terms of outcome measure. MRI was very helpful but just as a screening test, finding spinal cord abnormality issues. And pulmonary

function studies, well, they're practical only at age 7 or older. In our institution, we couldn't do them. And if you look at it, they're summative, you know? It averages the performance of both lungs. So it really doesn't tell you if a bad lung and a not so bad lung, what are each doing, so it's not too helpful.

Our results, we had out of 147 patients, it shut down 9 deaths. In the five-year survival by category, the bad actors were those children with very flail chests or small chests. The success rate, which is survival plus stable or improved assisted ventilation rating, was 83%, which was very favorable. The hypoplastic chest, they -- those that were on oxygen or worse, it varied from 48 to 83% before surgery, and 67 to 90% of those children improved. So it's not exactly a quality of life that's validated, but it's true clinical significance, and I think we've emphasized that earlier.

Our success based on thoracic enlargement purely on radiograph was only about 51%. As I showed you with the CT scans, which were hard to interpret with volume, it probably understates the improvement.

Adverse events. We had three cardiac arrests, 20 pneumonia episodes, - and this was over nearly a 12-year time span -- and respiratory distress in 14. The only significant device problem was device migration. And what you see here is a very slow process. The device goes into the rib gradually, and then there's extra bone around it. But, you know, that takes

three to five years to occur, and it's often asymptomatic. It's more a nuisance than anything else. Loosening of spinal hooks can also occur with other forms of instrumentation.

I'd been waiting 17 years for this approval letter, which I finally received in August of 2004. We were also cited in a 2005 Institute of Medicine report to Congress as a pediatric device development example. The lessons I learned, well, I'm often asked why did it take 17 years. Well, rare diseases are rare. It took a long time to get the numbers. I didn't know what I was doing half the time as a device developer. And I think it was also a learning experience for the FDA. And I really appreciate all the support they gave us through the years to help us get through this. What we both learned can help present and future innovators get their devices to market a lot sooner.

We need more advanced assessment measures -- this is a dynamic lung MRI. This is a child with neuromuscular scoliosis with lowered vital capacity. You see the kidney is pressed into the spleen and it is blocking the diaphragm. . And here's a normal child that we have scanned. You see the difference.

This is a spinal muscular atrophy patient who's clinically getting worse. And you see over 9 months of follow-up, you see how the thorax is collapsing. The spine doesn't look too bad, but the 3-dimensional thoracic problem is very bad.

This treatment is called a VEPTR gantry, expanding the chest. He clinically is a lot better. We measured this by calculating difference in volume during breathing of and the lung and also the diaphragm. This is a thoracic performance index, and it shows chest wall and diaphragm, and you see -- I'm almost finished. You can see thoracic performance natural history, you see very little chest wall motion pre-op, and then after treatment, there's an increase in performance here.

So you see the improvement with this advanced technique. We can't get money to fund this, to validate it, but we hope someday we can roll this out for you.

Animal models are difficult. This is a rabbit model by Dr. Brian Snyder at Boston Children's Hospital that shows changes that are favorable with a VEPTR-type approach. As far as quality of life instruments, it's probably helpful but not definitive because we think parents are stressed by their child's illnesses and shortened lifespan probably read a lot of themselves into their answers to the instrument. I don't think we should be too demanding about that. And I want to just say following these patients for many years, an illness without cure, a good year, a good month, or a good week for a child can be priceless.

We need better controls. We need to use our registry data, and you know, we need to do something like forensic analysis, like Dr. Brian Snyder advocates, and look at this data and clean it up. This is the concept of

adequate versus inadequate fixation and spinal surgery based on Dr. Behrooz Adbrania's work. And we can do this. We clinicians can help the FDA.

We need a universal IRB for these children. The CHOP IRB took over four years to approve it. There's not much device expertise in the IRBs, and maybe a federal IRB would be the answer for that.

My suggestions, in closing, we need to designate clinically relevant endpoints for innovative child's devices based on known natural history and a priori, with support of AAP and NORD, get expert panels together and look at both incidence and ranking importance of outcome measures; be flexible and pragmatic; the regulatory pathway has to be affordable and predictable; if you stop a trial, it may be a game-ending event; and really, if a gold standard is not available, why not silver or bronze?

Final slide, you know, life is a parachute ride. We hope the ride is long and it's enjoyable, but for some children, the ride is short. It can be meaningful if we really go to bat for them.

Thank you very much.

(Applause.)

DR. MIRZA: Great. Thank you, Dr. Hays [sic]. Now we have -- sorry -- Dr. Campbell, thank you. And now we have Dr. Ron Hays from UCLA's School of Public Health.

DR. HAYS: All right. I know we're at a rush for time. I just want to quickly say that Gordon Guyatt's presentation, I actually was very

impressed with it, and I was amazed at how much he's actually influenced me. I thought some of these ideas were mine, but I see they came from him. Gordon went through everything that I think we need to know about MIDs.

And what I'm going to try to do is just probably reflect some of the things he said with a concrete example if this works. Ah, there we go. Sorry. Where are we, here? Okay. I've lost all my time. There we go.

Okay. So what I'm going to focus on is physical function even though pain was in some older versions of the agenda and one example in the Patient-Reported Outcomes Measurement Information System project, which hopefully some of you have heard of, and I hope these are the current slides. This is one example of an item that's used to assess physical function, you know, whether someone's able to get in and out of bed. And I have a feeling these may be the old slides, so I'll do the best I can with them.

But we have a battery of about 124 items. And they are never administered in full, but that's a full bank just for testing purposes, and then there a lot of short forms. So I'm going to focus on a short form that has 20 questions that was administered. And actually, Jim Fries is the one who did this study, so I was just analyzing the data, so he can be the one that can be criticized for the design. But I used the data as best I could, and it's a pretty good study in some ways; there are some limitations.

Basically, he did an observational study of about 451 people with rheumatoid arthritis, and he assessed them at three time points,

baseline, 6 months and 12 months after baseline. And this 20-item physical function short form was administered at all three time points along with two legacy measures, the SF-36 physical function scale, which you've heard mentioned earlier today, and the HAQ, the Health Assessment Questionnaire. So this was done in PROMIS so that you could look at this new measure and compare it to existing measures that are widely used. And, of course, we're talking about self-reported physical function here.

The sample characteristics are shown, just to let you know, a majority were female and white, and the age range was 20 to 70-plus. And it actually didn't have the exact upper bound. It was categorical data, and it was presented that way to respondents. But the average age was 65, so it's an older population. And education was about 14 years, on average.

Now, this is where we get into what Gordon talked about earlier. In this particular study, at the two follow-up points after baseline, so you have the 6-month follow-up and the 12-month follow-up, there is an anchor item that is used to help you with estimation of the minimally important difference. And this is the particular item that was used. It's very similar to what Gordon talked about, where people tell you about how they are now relative to sometime before. In this case, they're asked about 6 months ago because the first follow-up is 6 months after baseline, and then when they're assessed again 12 months after baseline, they're asked about 12 months ago.

So you can see the item stem here, and it's basically focused on physical functioning, and there are five categories that are used to get at what they perceive their change to be, from "got a lot better" to "got a lot worse," and then in between has "stayed the same."

The better group is going to be the people on that retrospective item that say either they got a lot better or they got a little better. So you can think of -- in some analyses, you might want to lump those two together. And it turns out, it's about 56 people. It depends on what time point you look at, whether it's the 6-month or the 12-month follow-up. It's pretty similar, but it's not identical.

So that means the majority of people in the 451 that are in this longitudinal study are either going to be worse or they're going to be staying the same. It turns out these are the numbers that reported getting worse. So you have more people reporting they got worse than reported that they got better, but you've had a substantial number of people that said they stayed the same. It was about 250 people. So that's the distribution of the data.

Now, one thing you can do with this data is you can look at these groupings, just people who said they got better and people who said they got worse, and then say, okay, on the anchor item, that's what they said happened from baseline to 6 months and then from baseline to 12 months. So that's our anchor that says: Do they feel that they changed and in what direction? And then you can look at the physical function measure, and the

three in this particular study are the PROMIS measure, the SF-36, and the HAQ, and you can look at what actually happened prospectively on those measures. And in this case, we're reporting an effect size measure to say what actually looked like was happening in the prospective measure relative to the anchor.

So this shows in the PROMIS measure for those people who said they got better, they changed by an effect size of .21. And effect size is, you know, the change divided by the standard deviation, so .2 of a standard deviation change for those people who said that they got better on the anchor. And then the SF-36 and the HAQ are next to it, and there's the effect size that's associated with those two measures.

So they're pretty similar in terms of what they're saying about change if you use an effect size, a standardized unit, with the PROMIS measure in this case looking like it's reflecting a little bit bigger change, but not by much. And so this is Wave 3 versus Wave 1, so this is over a 12-month time period.

And then you can look at the people who said they got worse, you know, a little worse or a lot worse, and the same thing. And here, you see the SF-36, the PROMIS measure and the HAQ measure, and in this case, the SF-36 changed -- it looks like the effect size is a little bit bigger than what you see for PROMIS and HAQ. But, again, they're fairly similar in terms of the amount of change. If you use Cohen's effect size rules of thumb, these are

going to be in the small range for effect sizes. So when you lump together whether they got a little better or a lot better or you lump together a little worse or a lot worse, you're getting prospective change estimates in the small effect size range.

And if you look at Wave 3 versus Wave 2, which is another 6-month -- which is a 6-months difference, you can see a similar pattern of results. They're not identical, but basically, the three measures, the new measure, the PROMIS measure, and the two legacy measures seem to have relatively similar effect sizes when you're doing a relative comparison although you do have a very small effect size here for the SF-36 getting better. But, again, the SF-36 for the people who got worse, that looks like the effect size is a little bit bigger than the other two measures. And you have kind of a small change here for the HAQ. But, overall, at least the general direction seems to be consistent across the measures.

Okay. So this is pretty much showing the same thing. The only thing I've added here is that you can also compute an F statistic, which is sometimes referred to as a relative efficiency statistic, that summarizes across all the change groups and gives you an idea of how sensitive the measure is to that change across all the groups. So that includes the people who said they got better, the people who said they got worse, and then people who said they got the same.

So if you use that relative efficiency statistic, it actually looks

like, overall, the PROMIS might be a little bit more sensitive to the anchor in both cases regardless of which comparison you're doing, Wave 3 versus Wave 1, which is 12 months, or Wave 3 versus Wave 2, which is the 6 months difference in time. But still, all the measures seem to be sensitive to the anchor.

Okay. In this case, just so you can see, the same group is shown here. So the effect size for the same group is pretty much around 0, which is what you hope. So people who said they stayed the same over 12 months or over a 6-month time interval, their change is essentially 0 or a little bit more than 0, but it's definitely smaller than what you see for the other groups. And what's nice about this, as you can see, you hope that these are all positive, you hope that's close to 0, and you hope these are negative, and that's what's happening for every measure; the same thing here when you look at the Wave 3 versus Wave 2, which is the 6-month time interval.

So, overall, the results are suggesting that the PROMIS measure is either similar or, in some cases, maybe a little better in terms of sensitivity to change when you're using this particular anchor, this retrospective self-report anchor.

Okay. So that's what you always want to do as a start when you're going to estimate the MID. You want to make sure everything makes sense across the board. So when you look at the people who said they stayed the same versus got better or worse, does it look like you're getting the

monotonicity that you'd like, the relative magnitude of change prospectively.

So given that that's the case, that's suggesting your measure is responsive to change on the anchor that you're using, and you'd like to use more than one anchor; in this particular study, we only have this one. But once you've gotten to that point, then you can get to the point of estimating the MID.

So what we show here now is all five change groups on the anchor. And here on the raw score is the amount of change that occurred prospectively from over 12 months and then over a 6 months' time period. – In this case I'm just focused on the PROMIS physical function scale, because that's what we're really interested in estimating the MID for because it's a new measure. And it is scored on a T-score metric, so the mean is 50 and the standard deviation is 10 in the U.S. general population just like the SF-36 Version 2 is.

And what you see is what you'd like to see, just like you would suspect from the effect size estimates, is that people who said they got a lot better, that's the biggest positive change in both comparisons that are done here. People who said they got a little better, that's not quite as big. People who said they were the same, their change is close to 0. And then when you go in the other direction, there's a negative change. So people who said they got worse, they're getting worse on the prospective measure.

So once you have that, that's just confirming what you already

knew from looking at the effect size before when we were comparing all three measures. But what we want to focus on for the MID is this group, a little better, and that group, a little worse. And it's nice to look at them separately because, as people have said, it may not be the same. Going a decrement in health versus getting better in health, you might not get the same MID estimate.

It turns out we get fairly similar estimates, but what this would suggest is that the MID that's estimated from this anchor is somewhere around 2 on the T-score metric; I just sort of round it off. And if I'm looking in the other direction, it's somewhere around 1 to maybe 1.5, somewhere around there. So they're both pretty similar, but they're somewhere around 1.5 or 2, depending on which direction you're looking at.

So if we believe these estimates of the MID, given that the standard deviation is about 10, that's suggesting that in this case, the minimally important difference for the PROMIS physical function 20-item short form is somewhere around .2 of a standard deviation or somewhere around the small effect size.

So this is sort of just an illustration of what you're doing when you're estimating the MID. Of course, if you want to do it in reality, you wouldn't want to rely on just one anchor. You'd like to have some clinical anchors and as many anchors as you can get so you can see if you get convergence or not. But we didn't have it in this particular study.

So the summary is that we need to have anchors that are going to indicate change in physical function, and they have to be independent of our prospective measure, the measure we're trying to evaluate; in this case, the PROMIS physical functioning scale. And we have to realize that when we're estimating the MID, it's a sub-thing, a subpart of estimating sensitivity or responsiveness to change. And Gordon hinted at that. He didn't emphasize it a lot. But first we need to see that our measure is really responsive overall, and then we can estimate the MID, which is a subpart of that. In this case, it looks like the physical function measure was responsive to change, and it was similar or better than the legacy measures.

And it's important when you're estimating that MID that you focus on what the anchor is telling you is a small change but not a trivial change. And that's what's very difficult about it, is trying to figure out where is that range of change that's small but not trivial and not large. And that's the focus for the estimate of the MID.

Okay. And I'm just going to leave this up here for any e-mail correspondence. And this particular article I think is a good summary that people might want to look at if you haven't seen it already as well.

Thank you.

(Applause.)

DR. MIRZA: And our final speaker today is Dr. Traci Leong from Emory University.

DR. LEONG: Okay. I was going to say good morning, but I guess it's -- well, it's still good morning, right, or is it afternoon yet? Well, okay. Good morning. I'm going to talk about "Adjusting for Prognostic Variables," but you're going to notice that I'm really going to be touching upon a lot of the points that the previous presenters had made. And, hopefully, I'll be able to kind of pull a lot of those together.

Uh-oh. I'm not very good at this. Oh, there we go. Let me just start off with let's just define what a prognostic factor is just so that we're all on the same page. So the definition, you can see up top, is a patient characteristics that is related to a disease outcome regardless of treatment. And it's sometimes referred to as a risk factor, and sometimes they can be related to the disease like a staging system, and sometimes they can be related to one another. I gave the example of the Risser score in age and scoliosis. I'm not saying that they are prognostic factors, but those two variables are related to one another, and that was really the point I was trying to make.

My talk is really kind of piggy-backing a lot on what Aileen said earlier in terms of the patient factors. So as you remember from a few talks ago, she looked at the literature and looked to see what types of variables patient factors seemed to relate to different outcomes, and once those outcomes are known, then we can kind of proceed with the plan of attack that I am proposing here.

But before we get to that, I do want to show a slide that was a graph in the *New England Journal of Medicine* a few issues ago. And if you didn't happen to notice it, let me just explain that the bottom is chocolate consumption down here at the horizontal axis, and the outcome here is Nobel laureate. And so the question was, is chocolate consumption, is there something in chocolate that will increase the mental capacity and allow for more Nobel laureates per capita. Here, it's based on 10 million people.

And, of course, I'm not trying to tease you with chocolate, as we're about to break for lunch soon, but I do think that this point here is that you can see this nice linear relationship between as you increase your chocolate, what happens to the number of Nobel laureates. Well, they go up, too. Of course, Sweden, who hands out the Nobel Prize, they seem to be an outlier, but again, this has a nice linear relationship.

So is the answer that we all go out and buy more chocolate? No, because association is not causation. And so when we get to that point, our next step is to design and conduct a randomized clinically controlled trial to determine treatment effect and then which factors are most associated with outcome.

So I would say the point here is really how to frame the right question in the design phase. And I'll get to what happens after the trial has been conducted, but I really can't emphasize this point enough. Mark Melkerson mentioned sample size issues at premarket, and I'm going to

talk about that also, but here, the goal is to have both clinical and statistical significance. And unfortunately, and as some of my previous presenters have mentioned, clinical significance is not often agreed upon, and clinical success is not one that has been one of consensus. And the issue is to kind of determine the risk/benefit. As Dr. Campbell mentioned just now about the pediatric population, the risk/benefit among adults sometimes is viewed very differently than among the pediatric population. And so that is something that is important to quantify at the onset.

So when is it important to recognize prognostic factors? Well, first, there's the epidemiology portion of it that Dr. Tosi had mentioned earlier. And that refers to the incidence of the disease. And I'm going to mention that only briefly. And then I'm going to spend a little bit more time on study design as well as once a study has ended.

So just to have one slide on the epidemiology, the epidemiology in terms of these type of prognostic factors or risk factors are just basically looking at those characteristics that are associated with disease development. And Dr. Tosi talked nicely about how the female gender had a lot of those issues in several different orthopedic diseases.

Study design, I'm going to mention both an equivalence design as well as a stratified randomization. These two are not separate topics. They're not necessarily mutually exclusive, but prognostic factors can play a role in each of them.

So in an equivalence setting, the goal is the degree of equivalence between an old and a new product rather than superiority. Now, superiority, of course, is the gold standard, and that is how many of the trials and clinical research tend to be designed. Unfortunately, when you have -- well, it is not necessarily unfortunately -- it's fortunate for the patients -- when there are very high success rates, such as bracing in scoliosis there tends to be a very high success rate, and so in order to get a difference of, say, 75% versus 85%, you're going to need a lot of patients in order to determine that. And so when success rates are very high, one option is to design the study for -- based on equivalence so that we have a margin associated with non-inferiority instead.

Stratified randomization, what that is, is when we force a balance in the randomization process between the two different treatment groups based on known factors that will influence outcome.

Now, where the sample size comes in and where this whole idea of how do we bring our device to market a lot less expensively will come in will be here, because you can increase your power, meaning in the end, you can decrease your sample size, by having a stratified design. So having a stratified design versus a non-stratified design, you can decrease your sample size between 12 and 42%. Of course, this is all based on knowing that something is an actual prognostic factor. And once that is the case, you can design your study based on stratification in order to reduce your sample size.

Similarly, your power will go down and your sample size will go up if prognostic factors are unequally balanced by treatment and you did not stratify them in the randomization process.

So once we have framed the question, we've designed the study based on whether it's equivalent, whether you've decided to randomize through stratification or not, you've conducted your study, and now what? Well, the nice thing is that there are statistical models to adjust for confounders; there are statistical models that will help adjust for prognostic factors. But unfortunately, many times, the strategy is to use this to correct it on the backend, to correct the design deficiencies on the backend. And, unfortunately, that's not the most efficient use of a clinical trial.

So why do we adjust for prognostic factors? Well, the general purpose is just to learn more about the relationship between several predictor variables and a dependent variable. But as I said, imbalances in your prognostic factor will not only lead to an inefficient trial, but it could lead to false conclusions because you can improperly attribute an effect observed in the outcome variable to an intervention when it was merely due to the imbalance of the prognostic factors between the two groups.

So what is the alternative? Well, what happens, and Aileen had mentioned that in her talk, is that several times in publications, people only report a univariate analysis, so just looking at one variable against the outcome. When that happens, unfortunately, you're not simultaneously

controlling for the effects of other independent variables. And we also might not be quantifying the relationship between our outcome and our prognostic factor appropriately.

So typical analysis problems that aren't fixed by prognostic adjustments are listed on this slide, meaning that even if you can find prognostic factors, you decide to adjust for it, there are some things it cannot fix, such as if the true effects are weak and inconsistent, if your endpoints are poorly defined, if you are searching and searching for statistical significance kind of like a fishing expedition, and if you have important confounders that are being ignored.

So, in summary, when you don't adjust for a prognostic factor, and that is a prognostic factor that is known to affect outcome, you will end up with a less efficient trial, you might possibly get a false conclusion on the treatment effect, and if you do not adjust at the design phase, then you will be underpowered during data analysis.

And that's it.

(Applause.)

DR. MIRZA: So that was wonderful, and we are back on time almost. So we are going to have a short question and answer session. Deborah, if you want to -- will be moderating. So if you have any questions, there's mikes, and we have a runner on the floor to field questions.

DR. MOORE: Do we have any questions in the audience?

DR. MIRZA: You've stunned them into silence.

DR. MOORE: Yes. Okay.

MR. BOST: I'll make a quick comment, and this was regarding the first presentation where we saw the minimum clinically important difference being computed for blood pressure and for tumor response. And I'm not convinced that it is wise to compute MCIs or MCIDs for surrogate outcomes. I think that they probably play an overblown role already in modern medicine, and to imply that an MCID for a surrogate is something that a patient cares about may be just a bit much.

DR. SLOAN: You mean for the survival?

MR. BOST: For tumor response and for blood pressure. Tumor response is a very, very weakly if at all correlated to survival, for example.

DR. SLOAN: Right.

MR. BOST: Blood pressure probably not a substantial correlation.

DR. SLOAN: Sure.

MR. BOST: I'm wondering if --

DR. SLOAN: I guess I'm not clear. So you're saying that we shouldn't be --

MR. BOST: Well, I'm saying the MCID is a difference that matters to patients. And it's something of an oxymoron to take an outcome that does not matter to patients and then say there is a difference in it that

does.

DR. SLOAN: Well, the present definition of it is it's a difference that matters to anyone that might take action. You saw Gordon's updated definition. That's from, I think, his 2005 paper, right, that it's a difference that would cause -- that a patient thinks -- a patient or clinician thinks is important and that clinical action would be taken, so it could be either/or. But I mean, certainly, the context within which these effect sizes are drawn is important, you know? It's got to make sense within the context. I think that's what you're saying, right? And that's --

DR. HAYS: And I was just taking it as a kind of crazy example that shows the parallel, not that we necessarily would want to advocate doing that although there -- I mean, it is done implicitly, anyway. I mean a lot of times -- not blood pressure, per se, but you know, 50% reduction in seizure frequency or so forth without any basis. So some of the techniques that are used here pretty rigorously with patient-reported outcomes could be applied to clinical measures as well. You may not want to call it the same thing, but it's sort of a parallel idea.

Same thing is I thought Jeff was going to talk about there's a paper that Beth Hahn was first author on it, compared reliability of patient-reported outcomes with clinical measures, and it's a very useful paper because it shows, you know, they're often similar, and if anything, sometimes the patient-reported outcomes are more reliable. But you can measure them

in the same way.

DR. BOYAN: I was going to make the Dr. Phil microphone person come to me, but I'm over here now. Through the whole thing, I was really struck by the fact that I'm totally frustrated. And when I listen to this, how can you separate any of this from the context and from who the patient population is that you're trying to determine these things for? And for everything, it sounds like it's got to be different. So who sets the standard for what it needs to be, what it is in any particular application; particularly was the talks where we are discussing the reality of life, which is people are really different, so -- they're different ages, different diseases, different sexes. And I'm wondering how we're going to wade our way through all of this to give any kind of -- come to any kind of conclusion about the value of this number when it is so dependent upon this social -- or on the context in which it's determined.

DR. DAVIS: So I don't disagree with what you've said, but I think those of us who would be the math junkies around measurement would argue that measures -- numbers on a measure themself even don't have reliability. The reliability of the measure is context-specific. So what we're saying about the MCID and the fact that there is some level of context-specificity to that number is just what we say about all these measures.

DR. SLOAN: And that's where I tried to emphasize the message is that it's good news even though you can get lost in the weeds, as a

colleague of mine always puts it, in terms of small differences and that sort of thing. For example, in comparing the discussion between your talk and your talk, you're talking about the difference -- on one level, you could say, wow, they're saying different things. But really, they weren't saying different things. I think both went to great lengths to say that -- is there's definitely gender differences, and it definitely needs to be looked at and it's important, but also, in terms of the size of the impact on the variables that we're looking at can be quantified and incorporated into an analysis so that you can take that into account.

For example, very rarely are things like age, gender, and education accounting for more than 10% of the variance in something that you're looking at. That doesn't mean you ignore it. But it means that it's within context of how different it is. And I think both speakers gave excellent examples of where the evidence was, I think it would be fair to say, questionable as to where it was on the fence, and it required a real careful exposition and interpretation of what the data are saying.

And that's okay. And so you can say -- I think -- to me, it's the difference between the statistical and the medical. Statisticians look for the signal amongst the noise, amongst all the variability. Medical professionals, by very nature of practice, have to look at the -- make sense of all the noise in terms of one little -- if you watch the series *House*, right, there's always one little bit of data that is unique to that individual that causes the whole thing

to unfold. And so that's the way medical professionals tend to look at things, where individual bits of data -- so you're coming at it from a different perspective.

And I think that's where -- from -- in my experience, anyway, the medical folks in particular say, wait a minute, what about this and this and this. And you have the statistical say it's okay, it's all right. Well, have those things, but within the "this" and "this" and "this," there's a big "that" that is the basic message. And that's where when we come together in that perspective, I think that's when we can come to the consensus that you're talking about. We kind of have to recognize that different people have different -- it's almost like a style sort of way of approaching science.

DR. LEONG: And I don't think we can underestimate the issue of the regulatory situation between pediatrics and adults. And I think that Dr. Campbell mentioned that also, and I'm in full agreement with IRB issues, with pediatrics and devices, you know, what is acceptable risks for pediatrics, very different, and it's not standardized, it's not something that people have a consensus on. And I think that also -- kind of on the opposite end of this data analysis and interpretation aspect is I think an area that needs work.

DR. HAYS: I wanted to say also that you'd like to look at whatever groups you think where there could be a difference like gender, age, or whatever. A lot of times, you just don't have the data, and so it's not done; you don't have enough people, so you barely have enough in these

change groups in the overall sample. So if you have the resources, you'll look at whatever you can that's plausible.

But I think in the end, you know, you're going to find some cases where it's going to vary and maybe systematically, but the overall message is probably going to be that it's kind of similar, and we may have a few exceptions that we need to take into account. And we never want to dismiss them, but you know, I don't think we're going to see huge differences in what's important to people. You know, we don't know in every case, and whenever we can evaluate it, we do if we have the data, and we should.

DR. CAMPBELL: I wanted to add, Barbara, I quite agree, you know, this is just bewilderingly complex, and it just -- you know, you just don't know which to turn -- I think, though, we have to acknowledge, it's not impossible; it's just damn difficult. And I was talking to Mark Melkerson earlier today, and he was -- he mentioned -- he said, well, we've been trying to do something like this for 25 years. Well, we've finally done it. We're talking about this. And the first step is define our problems, and we're getting a better handle on that just with this morning, and then work on solutions. You know, the kiddie side is very difficult, but we're making progress. You know, we're not doing spine fusions to one-year-olds like we were when I started in orthopedics, and I think that's because we learned something. Now we have to learn to measure it.

DR. BOYAN: It's on, okay. I guess I wanted to hear a little bit

from Debbie, because I happen to know that Debbie is interested in measuring pain, and that's something that really is a much more -- it's not an easy thing to measure. It's a much more difficult thing to measure and very, very person -- it's person-specific. It's not like large groups. And here's a case where someone has to make a decision about whether or not it is medically important or not -- minimally important -- whatever we're calling these things now -- whether it matters or not.

And at some point, there is a number that has been somehow collectively arrived at that is believed to be meaningful. And I guess that, in my mind, is what we're trying to get to in this conference is an assessment of what that number ought to look like. And where are we going to go to get that piece of information that we can then all agree upon is the magic number.

Deb, did you want to make any comment on that?

DR. MOORE: Sure. I'll comment on that, as well as I had a question for the panel regarding that particular topic, which is I think on the pain literature, there's a lot of different literature that looked at the clinically significant importance of pain; I think it was the IMMPACT group that had a consensus that looked at the literature and came up with about a 30 to 33% change in pain was determined to be clinically significant, and what was the recommendation for when you're evaluating clinical trials that are looking at pain, you look at that particular difference, and you would consider that

relative change. So you're not looking at an absolute number. And I think that would take into account, perhaps, maybe some of the differences that you're seeing with women if they're coming in with greater pain versus a male that's coming in with less. I mean, I think if you used a relative change, that might kind of even out that effect.

But I guess the other issue that I wanted to ask the panel is, I mean, in addition to looking at just the MID, it's what MID are we going to be looking at? Is pain more important? Is function or quality of life? I think there's a lot of different tools that are out there, but as a person sort of working with the Agency, we need to kind of come up with what is our primary analysis going to be that's going to define success. And what are the important tools? Is one of those measures more important that you've seen that has the better predictor of what's the most important to a patient?

DR. DAVIS: I'll speak to the latter in the context of joint replacement because it's what I know best. So one of the other challenges we've got for a lot of these orthopedic conditions is when you look at the patient population, we're dealing with either other comorbidities or, in the case of arthritis, that this is not a single joint disease; it's a multi-joint disease.

So for something I was doing, I had some colleagues pull the data from the Swedish joint registry. And so one of the challenges you've got is about 30% of people who have had a single hip or knee replaced go on to have another hip or knee replaced. So we, in fact, in some of the work we've

done in our follow-up in trying to understand some stuff, we've actually taken out those people who had their second -- had a second joint replaced, because then if you go to something like quality of life, is it about the device that you've put into that one knee or now are you starting to get into these other issues.

So part of that is when do you do your follow-up and for how long and what those -- what else is going on in that patient's life in terms of health issues and what have you. So it's challenging to know what that is.

The other part, I guess, from my clinical hat is that -- and one of the speakers alluded to it -- it might actually -- I guess it was Gordon Guyatt's talk -- looking at these separate domains. We see them change at different rates in a trajectory of recovery, potentially. So if you decide that, you know, in a drug trial it's at 3 months, well, if fatigue might have been one of the major factors, fatigue may not have recovered by 3 months, but by 6 or 9, it might have. So I'm not saying there's a perfect answer, but I think those are things that also need to be thought through carefully.

DR. SLOAN: The other thing about pain is that it is an example where there's a vast body of literature defining, you know, what's an important level of pain, and so on. You know, there's a paper by Tito Mendoza in Charlie's group, Cleeland, at MD Anderson years ago talking about what's mild, moderate, and severe pain and, you know, looking at data from a long period showing this is what seems to be not clinically important,

clinically important but not urgent, and urgent, and classifying it like that for that particular purpose.

And is it perfect? No. But Charlie's actually got a series of papers coming out looking at different things like that, where we've used, for example, clinically, for overall quality of life and fatigue on a scale from 0 to 10, if you have a 5 or lower, we've related that to, okay, 50% increase -- or a doubling of the risk of death in cancer patients, okay? So that's the clinical linkage, the anchoring, again.

Charlie's looked at, well, okay, you've got 5. What if it's 4? What if it's 6? You know, that's the classic question. What if it's something different in different contexts? And we've looked at it in all different ways. And you know what? In some certain circumstances, it might be 4; in some circumstances, it might be 6. But as Ron showed in his data, it's marginally different; it's just a little bit different. And so you won't get a totally wrong answer -- you'll get a different numerical answer, but qualitatively, the question as to whether or not your device is useful or not is not going to change because you've used a particular cutoff. And, again, it's the idea of doing sensitivity analyses, let's use this other cutoff instead, and that's the beauty of computers.

MS. STONE: Thank you. We've have another question from the website. Dr. Moshe Vardi, Associate Medical Director of Harvard Clinical Research Institute asks: How would one account for a patient perception on

the intervention when looking at patient-reported outcomes? As an example, how would you assess a functional score of health-related quality of life scores in a trial assessing surgical versus non-surgical intervention?

DR. SLOAN: Well, one of the things that we've done more recently is come up with a tool called a "Was It Worth It" measure where we ask patients, you know -- this is tougher in pediatrics, of course, but a very simple tool, ask patients three questions: If you had a choice to do it over again, would you do it? Was it worth it to you? Would you recommend it to others?

You know, and interestingly enough, when we first developed the tool, people were convinced that, well, if you get a positive clinical outcome, patients are going to think it's worthwhile, and we just presented data this year to demonstrate that that's actually not the fact. It depends how they're treated, and it goes back to, I think, to Dr. Campbell, you know, when you're saying the QOL data didn't quite match what your clinical experience was, I'm guessing if you asked your patients, you know, that didn't see differences in QOL whether it was worth it, I got a sense that many of them would have said my QOL data may not have changed, but it was worth it to me. Fair statement?

DR. CAMPBELL: Let me mention something. One of my earlier patients died at age 23. I went to the funeral. His parents were divorced at that time, so I went to them separately. And my big question was exactly

that. Was it worth it to your son? And both parents independently told me -- they said, you know, we -- he went through a lot, multiple surgeries, a lot of hospitalizations, but they said, yes, it was. He wanted a chance. He got it. And he had well-rounded, rich 23 years. I wish it had been more, but that was the answer.

DR. MIRZA: One last question.

AUDIENCE MEMBER: We've talked a lot about efficacy outcomes and the minimal difference for efficacy. I wonder about safety. If you had an equivalence design between two therapies, and there was a difference in safety, is there -- could someone tackle the thought of a minimal important safety difference?

DR. DAVIS: So sort of safety, but I guess effectiveness, I think the hip resurfacings that became popular sort of in the last 5 to 10 years again are a classic example of that. The bottom line is the devices overall initially were being put in quite broadly across younger people and what have you, but the bottom line is the device failed. So if you looked at the initial short-term data, patients were getting great pain relief, function was improving and what have you, but now they've actually -- because of failures and the device itself, they've now honed down to very -- from my surgical colleagues a very small subgroup of people, which tends to be males sort of in that 45 to 55 range who still want to be relatively active, and they're staving off total hip replacement.

So in that situation, to me, the failure of the device or what I would call equivalent to what your safety question is out-trumps the patient-reported outcomes in the short-term. But that's my opinion.

DR. HAYS: I'd also say I think you could evaluate it the same way, though. It's just you might be looking in the other direction. So you could say It's the MID for getting worse. And, you know, if it's getting worse beyond the MID, it would be something of a concern.

DR. MIRZA: Great. Thanks, everyone. That was a wonderful discussion, and I'll know we'll have a lot more questions, and we can continue in the afternoon. And then we'll have great stimulus for a discussion for the workshop breakout sessions tomorrow.

We are a little bit late, but if we can reconvene back at 1:15, I would appreciate it. And have a good lunch.

(Whereupon, a lunch recess was taken.)

A F T E R N O O N S E S S I O N

(1:17 p.m.)

DR. MIRZA: All right. Welcome back, everyone. We're going to go ahead and get started just to try and stay on time. And we may actually finish early this afternoon because we actually had one of our speakers, Dr. Obremskey, he missed his flight connection, so he will be unable to attend

and give his talk on lower extremity. And Dr. Sedrakyan came down with a fever and is ill and will not be able to give his talk. So we may actually combine the two, skip through the break, and finish early if that's okay.

Now, one issue I did want to just bring -- to try and bring together what the FDA is looking for -- and a number of questions were raised in terms of, you know, what is the purpose, we're talking about all this statistical jargon, meaningful benefit, how does it all tie in. And I essentially wanted to summarize the purpose of the critical path project for which funding was received for this project, of which part this workshop is one deliverable for the first year of this project, is to essentially determine the minimum clinically important difference and how it impacts science and regulation and see if this is the metric of outcome and how it can be impacted by different variables.

So, for example, if you have a device submission that has a particular feature that requires clinical data or it is a PMA or de novo, then in that particular case, they may provide some clinical outcome measure. And then to understand what the minimum clinically important difference towards -- as Mark mentioned before -- towards the sample size power and, ultimately, the composite success criteria and study success criteria, all of that is relevant.

And understanding how that MCID varies for different target populations which may be different in women, men, gender, lower

socioeconomic status, BMIs, all of those may factor in to determine, well, is this an appropriate MCID in your calculation. And whether it matters a little bit or a lot, I think, is just as important as knowing where the FDA is going with trying to gather the information from all of you experts and seeing how we can use this to improve our understanding of device science and regulation.

So without further ado, I would like to introduce Dr. Marc Hochberg. He's a renowned rheumatologist at the University of Maryland Medical Center. And he promised he will keep his 20 minutes on time.

DR. HOCHBERG: Thank you. Oh, I should have gotten -- wait a minute. I needed five minutes of intro to figure out how to work this.

Okay. So I'm going to talk to the abbreviated group -- okay. Very good. It's always tough to be the first speaker after lunch because either you have a much reduced audience because people don't come back on time or, two, people fall asleep because of the post-meal sort of doldrums in the afternoon.

So I'm going to give you the rheumatologist's perspective on minimal clinically important difference or minimal clinically important improvement, okay? So the other thing is when you speak after lunch in the afternoon, you have no idea what's going to come in the morning, so some of this will be repetitive, and I'll try and skip over that quickly.

So you've already heard about the amount of improvement and whether this is measured as an absolute amount or a percentage change. We're talking about devices here, and in my experience, it's mostly been devices for treatment of osteoarthritis. So I'm going to talk a little bit about osteoarthritis trials. So it's either the improvement in the treatment group or the difference between the treatment and the control group.

Now, you've heard about effect size this morning, which is this sort of unit-free or scale-free measure of the relative size of the effect of the intervention. And you've heard about how to calculate the effect size. And you've heard reference to Cohen's classification of effect sizes, be they small, medium, or large, and one of the speakers focusing on the fact that, well, maybe a medium effect size or a half a standard deviation is something which is clinically significant, okay?

And, you know, we've recently conducted a large observational study, NIH-funded, where we were powered in order to detect medium-sized differences between two groups. The two groups were men and women. And we felt that when we wrote the grant for the study, that if we only detected small differences, that that really probably was not clinically meaningful, so therefore, we powered our study for effect size of about .5.

But in terms of the interventions that we use to treat patients at least from the rheumatologist's standpoint, most of these do not have an effect size when measured as the difference between the treatment and the

control groups in the trials divided by the pooled standard deviation, which reaches .5. So we're dealing in a world of treatments which have a small to moderate effect as opposed to a medium to large effect.

Okay. I'm pushing the right -- there, we're good. So now I'm going to borrow from an editorial which was written by Maxime Dougados in 2005. Maxime is currently the president of EULAR, which is the European League of Associations of Rheumatology, and he is a clinical investigator in Paris. And he has broad interests across many different diseases.

So he wrote an editorial where the first part of the title was "It's Good to be Better." And he talked about the minimal clinically important improvement and the minimal clinically important difference, and these were the definitions that he cited. And we'll stick with MCID, which was the smallest difference in the score of an outcome variable. Here, you have the three different domains which were referred to this morning, pain, physical function, quality of life, which patients perceive as beneficial, okay?

But he concluded his editorial title that "It's Better to be Good." Most people would agree with this; it's good to be better, but it's better to be good, because we know that people who start with a pain score of 9 on a scale of naught to 10, if they get better and they get to a 6, they're still not satisfied where they are at a 6. So they're not good.

And we haven't discussed at all this morning this concept of the patient-acceptable symptom state. I'm going to come back to it a little bit

when I talk about outcome measures, but this may be something you want to bring up during the discussion period.

Okay. So let's talk about this from a rheumatologist's point of view and focus on osteoarthritis, which is what I've done for the last 30 years.

So there are multiple ways of measuring outcomes in osteoarthritis. Dr. Davis will be pleased that I listed on here the OARSI-OMERACT osteoarthritis pain measure, or the ICOAP, but we'll start with the visual analogue scale or numerical rating scale, which is still reliable, valid, and a responsive way to measure pain and is used in a number of studies.

But mostly since the late 1980s, we focused in OA on using the WOMAC, or the Western Ontario McMaster Osteoarthritis Index, which is a 24-item instrument which either can be completed with a five-point Likert, or Likert, scale, depending on your pronunciation, or a visual analogue scale, 0 to 100, or a numerical rating scale, 0 to 10. And it has three different domains: Pain, physical function, and stiffness. Most people use pain and physical function separately. Some groups will use the total score. Nobody uses stiffness individually.

Then you have more recently developed the Knee and Hip Osteoarthritis Outcome Score developed by Ewa Roos and colleagues in Sweden; Ewa is now in Demark. And these are abbreviated as the KOOS and the HOOS. And these are freely available on the Internet. They're translated into a number of different languages, and they can be downloaded and used

without having to pay a royalty to anybody in clinical studies.

Then you have the more recently developed OARSI-OMERACT Osteoarthritis Pain Measure or the ICOAP, which stands for Intermittent or Constant OA Pain. And Dr. Davis was involved in the development of this along with Gillian Hawker and others from both of these organizations.

If we go back in time, we can still use the Lequesne Algofunctional Index, which is a single measure which aggregates pain and function, so it's fallen out of favor. Developed by Michel Lequesne and published in the French in the 1980s and then subsequently translated. This actually is useful because it has a point scale and an indication for total joint replacement based on the number of points that are achieved by the subject in the study.

And then, finally, you have the AUSCAN, which is used in hand osteoarthritis.

Okay. So what measures can be used and what are the outcomes? So for osteoarthritis, we have the Osteoarthritis Research Society was selected to respond to the FDA RFP on providing help with the revision of the guidance document for studies in osteoarthritis. As part of this, Bob Dworkin at the University of Rochester led a group looking at symptomatic outcome measures for trials. This would apply to devices as well as other interventions.

So he reviewed double-blind, placebo-controlled randomized

trials in patients with hip or knee osteoarthritis published before August of 2009. And 125 studies provided data on one or more outcome domains, and these were the three domains which were felt to be important. And these are the three domains that are used as co-primary endpoints for studies of pharmacologic interventions to get the indication for the treatment of osteoarthritis: pain, functional limitation, and patient global assessment.

In addition, there are other secondary outcomes. So categorical outcomes, as was mentioned earlier this morning, are considered to be secondary outcomes after you have demonstrated statistically significant differences in mean changes between the two treatment groups, okay? That's currently the regulatory stance at least with regard to drugs.

So you have the OARSI-OMERACT Responder Index, the minimal clinically important difference or improvement, the PASS, and then what was discussed during the discussion period before lunch were the IMMPACT recommendations of a certain percentage improvement in pain, either moderate improvement, which is a 30% decline, or substantial improvement, which is a 50% decline.

So this is what the OMERACT-OARSI Responder Index is. This is a valid measure where you can be a responder based either on high improvement in pain or function or modest improvement in pain or function or patient global assessment, two of those three things. This is published in 2003.

Because of time, I'm going to rapidly go through the next set of slides, which are data from one of our NIH-funded randomized trials in patients with knee osteoarthritis where we validated the OMERACT-OARSI Responder Index. So this was a three-arm randomized trial of 570 patients who were randomized to either traditional Chinese acupuncture -- some might consider this a device, okay -- sham acupuncture, or an education attention control group. So out of these 570, 41% achieved the OMERACT-OARSI Responder criteria at the end of the 24-week study. And this was 61% of the 386 completers.

So as you can see, the responders, who are shown on blue, had lower mean pain scores at the end of the 24 weeks than the non-responders. Well, that's sort of circular reasoning because that's how you define the response. But they also had lower pain scores on the health assessment questionnaire or on the HAQ Disability Index, which you saw was used as an anchor for the PROMIS measure earlier this morning. They had higher scores on all of the domains of the SF-36, not only the physical domains, but also the mental health vitality, et cetera, domains. And they had higher quality of life scores based on the EQ-5D or the Visual Analogue Scale feeling thermometer. So being an OMERACT-OARSI responder, it's better to be good.

Now, how does that compare with minimal clinically important improvement or difference? So we actually have an anchor-based study in osteoarthritis conducted by Florence Tubach and colleagues published in the

Annals of the Rheumatic Diseases in 2005. And here, this was the smallest change in the measurement that signified an important improvement in the patient's symptoms.

So these are patients with osteoarthritis to the knee, all of whom are given a non-steroidal, anti-inflammatory drug. At the end of one month, the patients were asked if they were improved, unchanged, or worse. And if they were improved, were they minimally improved, moderately improved, or markedly improved. So they got rid of the markedly improved patients, and they looked at the minimal and the moderately improved. And this was the absolute change in the Visual Analogue Scale from or to 100 for knee pain in knee OA and hip pain in hip OA in terms of the 75th percentile of the distribution of change scores.

So if you scored 20 mm improvement or greater, you've reached the minimal clinically important improvement, 15 mm for hip OA. And they also identified here the patient-acceptable symptom state, or if you reach this level or below in terms of your endpoint VAS at the end of the study, you were then in a patient-acceptable symptom state, 30 mm for knee OA, 35 for hip OA.

And when you look at the proportion of people in the acupuncture trial who fulfilled the MCII or PASS, you can see that it was about 30% for pain, 40% for physical function, and about 25% for global, and they were highly significant associations, and I'm not showing you the odds ratios

and the confidence intervals between having an MCII on one domain and having an MCII on the other domains. And the same was true with reaching a patient-acceptable symptom state, about 40% for pain and function and about a third for global. And also highly significant associations between achieving an OMERACT-OARSI response as well as having the MCII or the PASS.

Now, I mentioned the IMMPACT recommendations already.

And so we've analyzed all of these in the duloxetine trials. Now, duloxetine you know is a centrally acting agent which is FDA-approved for the treatment of major depressive disorders, generalized anxiety disorders, fibromyalgia, and more recently chronic pain associated with knee OA or low back pain.

And this shows you the summary of the outcome. So in the top row is the OMERACT-OARSI responders, and you can see that there's about 33% greater chance of having a response if you're allocated to active drug versus placebo, although note that 50% of the patients in these trials who were randomized to receive placebo achieved an OMERACT-OARSI response.

You can then see going down the proportion of the people who reached an MCII for pain and function, the proportion who achieved a PASS for pain or function, and the proportion who achieved the IMMPACT moderate improvement or substantial improvement. And the relative rates comparing active drug to placebo are all similar with overlapping confidence

intervals. And what these categorical outcomes provide you the opportunity to do is calculate the numbers needed to treat, which I think is important when you present these kinds of data to practicing physicians who really have difficulty interpreting what does it mean that there's a 3 mm difference between active treatment and placebo, as we see with a number of approved devices for treatment of knee osteoarthritis.

And you can also compare these NNTs to the number needed to harm, okay? And if the number needed to harm is smaller than the number needed to treat, well, that might not be something you would want to recommend to your patients. And that's the case for opioid analgesics in the management of knee pain in patients with osteoarthritis; the NNTH, the numbers needed to treat to harm, is lower than the number needed to treat to get a clinically important improvement.

So let me finish with some comments about the REFLECT study. This is a new study -- it's an old study, actually, but it's newly published, so it'll be new to you probably. So this is in press in *Arthritis Care and Research*. This is an international study which was designed to estimate the MCII or MCID and the PASS for four generic outcome measures in five rheumatic diseases in seven countries. And the five rheumatic diseases included rheumatoid arthritis, osteoarthritis of the lower extremity joints, osteoarthritis of the hand, chronic low back pain, Oxford Classes I and II, and ankylosing spondylitis.

We're scrolling through all the hidden slides now. Let's see. Okay. So this basically gives you the descriptive information on the studies or on the patient populations.

So this was the external anchor question for the MCII. Again, patients were asked whether they'd had a change from baseline on a three-point scale. If they reported improvement, they were asked how important the improvement was to them. And if it was slightly important or moderately important, they were then included in the analysis for the MCII or MCID. And the same question for the PASS here.

So 1500 people were enrolled in this study, 98% completed, 44.5% reported slight or moderate improvement, and 67% reported being in an acceptable state.

So I'm just going to show you the data for hip or knee osteoarthritis, again, because we're looking at here orthopedic devices, and the vast majority of those devices for a rheumatologist are implanted or used in people with osteoarthritis.

So 353 patients had hip or knee OA, and 249 patients had hand OA. So these are all on a scale now -- these are on a scale of naught to 100. So the absolute change was about 10 units, with a 95% confidence interval ranging from 6 to 12, okay? The relative percent change was 17%, with again a range running from 12 to 21 for the confidence interval. And the PASS was 39, 95% confidence interval shown there for pain. And for function, the

scores were a bit smaller. The MCID was 6, the relative change percentile-wise was 12, and the PASS was 48.

And if you look at hand OA, the results for hand OA are shown here, similar in magnitude to what was seen for hip or knee OA with a different instrument but the same underlying disease.

Now, in limited analyses in the original paper, there were no differences by country across the different diseases, and to my recollection, there were no differences by gender.

So I'll sum up here, stay on time, and thank you very much for your time and attention.

(Applause.)

DR. MIRZA: And thank you very much, Dr. Hochberg. That was a great talk. We're going to have questions and answers afterwards.

And I actually did forget to introduce Dr. Lynne Jones. She's going to be our moderator in addressing the question/answer panel.

And so I'd like to next introduce Dr. Charles Day. He is from Harvard University. I actually met him last fall when I was first putting together the proposal for this research collaboration, and he just was about to go off on a worldwide tour for his hand fellowship. So I'm sure he has a lot to share with us.

DR. DAY: Getting a special tutorial, here. Thank you, Faisal, for putting this together, and it's a privilege for me to be here to really speak on

upper extremity outcome instruments.

When I was thinking about this topic, actually, I was kind of -- when Faisal said -- gave me that topic, I said, are you kidding me? Fifteen minutes and you want me to talk about all of hand/upper extremity -- you know, from shoulder all the way down to hand? I mean, there's at least 50. So he toned it down. He said, no, just your experience. And that was a little bit more manageable.

So what I want to do is sort of go over my own personal perspective from a clinician's perspective. Ten years ago, I finished my fellowship. I did a lot of basic science research, did some clinical research. And as I started out embarking on my academic journey over the last 10 years, how did I utilize these instruments? That's how I want to take you guys through it and then my own personal thought processes as I encounter each instrument as I was thinking about each research topic.

And I'm going to talk about the DASH, going to talk about the Patient-Rated Wrist Evaluation, the Michigan Hand Questionnaire, and also the Modern Activity Subjective Survey of 2007, the MASS07.

And then we'll then sort of digress a little bit and talk about what I've learned over this past year. And this is an opportunity I've received from the American Society for Surgery of Hand, to spend a year sort of learning on different things. And I spent a lot of time at Oxford and looking at some of the patient-reported outcome instruments that they've developed.

And give you a report on what I thought were their take-home messages on what they think were the key messages and key components for them. And then how do our -- the ones that I've talked about compare to their -- what they recommended.

So the DASH actually -- I didn't know much about the DASH. I heard about this thing when I was a fellow. And as a fellow, we're all required to do a clinical research project, and I was doing a fellowship with Richard Gelberman, who's a very academic guy, and so I had to do a research project. And so immediately, he said you got to use the DASH; you got to get to know the DASH, and so I didn't know much about it. But at that point, I actually did a lot of this looking up after I started using it.

So one of the key things about the DASH is it's got a lot of acceptance throughout the orthopedic community. So even without a true understanding of it, a lot of people recognize the DASH. And to me, that is an important aspect of these PROs that we're talking about. It was developed in 1996, and more importantly, it was developed through collaboration. And that perhaps is the main reason why it is so widely accepted in the upper extremity realm.

It was sponsored by the AOS. It was sponsored by the upper extremity collaborative groups throughout North America; there's 10 different centers that went into this. And it was also sponsored by the Institute of Work and Health.

So these organizations putting together and sponsoring this allowed -- I think gave it the purview that it generated. The items were generated from literature review, clinicians, surgeon and expert input. Patients were only involved in the validity testing. They were not involved in the item generation.

It was then validated through the IAWH through a prospective trial of 109 patients, and there are some experts here in this room who can probably comment more on the DASH as it was being developed.

But just some key components of validation. I actually had to learn some of this myself when I had to validate one of the instruments I developed, but internal consistency; reliability; test/retest; validity, does it measure what it's supposed to, content validity as well as construct validity; as well as sensitivity to change. These are some of the key things that -- in terms of validating a outcome instrument.

Now, the DASH consists of two sections, about 30 questions. Its overall concept was the measurement of the overall upper extremity health, not a specific region of the upper extremity. And it's got two domains. There's a symptomatic domain looking at these things, and there's a functional domain looking at physical, social, and psychological function.

The scores are really calculated and normalized to between 0 and 100, where 0 is really no limitations whatsoever. And the lower score is better in terms of functional outcomes and symptoms. Now, I did look up the

MCID. There are some studies that have done MCIDs for the DASH, and it's recognized as 10 to 12 depending on what study you're looking at. But these were developed based on shoulder impingement and carpal tunnel surgery results. So I think it's also important to recognize that as people are -- some of these studies that are developing MCIDs, they're developed, the MCIDs are developed for specific conditions that they were looking at.

So I think my own personal positive experience is, number one, this is a multicentered and non-center-specific development and testing. So it doesn't belong to one institution, doesn't belong to one person. DASH is really sort of ubiquitous, and there's a whole --

So this is one of the big pluses of the DASH. It's an overall assessment of the upper extremity health. It's a good measure of any upper extremity problems.

I can see how this is -- some of the drawbacks. In the upper extremity, I can't believe that it didn't look at right-handed or left-handed, I mean, your hand dominance. I mean, when you're asking these questions, my patients are always asking me, well, should I answer with the hand I actually do it with or the hand I don't. And it also, it's not wrist-specific, and that's unique to me because my area of interest, as you'll see throughout my talk, is in the wrist. Its patient evaluation only really excludes patient -- surgeon input, and that can be a plus or a minus. And, again, no hand dominance. And it's relatively long.

Now, the DASH actually did address the relatively long part. Now, it's a 30-question survey. And to address that, the quick DASH was developed through the same working group to address the relatively long DASH questionnaire, and they were able to cut it down from 30 questions to 11 questions, and it's validated through the NIH to -- it's got strong correlation between the quick DASH as well as the full DASH.

So this is the study I was talking about where I was a fellow. I used this initially with basal joint arthritis of the thumb and looking at how effective it was in terms of after a steroid injection and seeing how the patients responded to that steroid injection. We did find a difference there.

And then when I started my clinical practice in Boston, that's when I started -- when you're a young hand surgeon in Boston, in a city where there's a hand surgeon every other block, the only thing left to me was trauma. So I did a lot of my research on trauma and distal radius fractures.

So this is one of my earlier studies looking at a particular type of plating technique for distal radius fracture. I used the DASH there. And this is another study looking at that technique. We used it there as well.

So as I alluded to, distal radius fractures is a big -- certainly, I had a huge clinical volume to do this when I started in practice. And it's also developed into a large clinical research area of mine.

And in addition to those plating techniques, I've also looked at other plating techniques. So here's a paper, it's a prospective randomized

trial looking at basically two different surgical techniques, volar plating -- so we talked about dorsal plating in one of the other -- this is volar plating and percutaneous pinning, so we've utilized it in these as well.

But then I realized a few years into my practice, I started seeing these patients that were very successful early on, in the first year or two. They started coming back and having some problems. And that's when I started looking into, well, are we having some complications, and some of the people alluded to some of this earlier on. So patient-reported outcomes are great until you start having complications down the line. And how does that fit into the whole picture?

So here, we did a paper really looking at two specific complications. So this is a four-year follow-up of complications in my own hospital of the different techniques that I initially looked at earlier on looking at what our complications were from fixation. And what I realized is that -- is this question. Complications are worth the risk if there's improvement in functional outcome. We know in distal radius fractures that the literature supports that anatomic fixation is important. But, actually, the older population, which happens to be the population that has the highest epidemiology of distal radius fractures, this is not clear.

Here is a paper by McQueen, 1988, "We conclude that the malunion of a Colles' fracture, or a distal radius fracture, results in a weak, deformed, stiff, and probably painful wrist." Another paper, 1989, "Strong

correlation between the functional outcome in both the dorsal angle and radial length at union." Just a few years later, 1991, "In patients over the age of 55, we found no correlation between the final anatomic and functional outcome." The radiographic outcome in greater than 60-year-olds did not correlate with the functional outcome.

So as I was thinking about distal radius fractures and where I was going to go, I realized then at this point, I know that fixation makes the x-rays look better. I know that there's some complications associated with that. The question is why is there such a differential in terms of functional outcome? Why is the literature so controversial in this? And my thought at that time was, well, they must be using different instruments or the instruments are not detailed enough or something. I have to find more than just a DASH. I've been using the DASH. And that's generally accepted, but it's not giving me -- it's not giving people the difference that they need.

So I started looking around for other outcome instruments. So is the difference based on outcome instruments? So we actually looked at this -- we actually then embarked on this paper where we actually looked directly at the relationship between anatomic reduction, radiographic outcome, and functional outcome. And my hypothesis was that I wanted to use as many functional outcome parameters I could to see if we can start comparing apples and apples and oranges and oranges.

So I started looking around; well, what's out there. Here's the

Patient-Rated Wrist Evaluation. It's utilized specifically for wrist problems. So this is something much more specific than the DASH. It was developed in 1998, and it seems -- there seems to be a pattern. All these things are developed in Canada.

It was developed in 1998, surveyed 100 international wrist investigators to generate items; 66 of them responded. And the item-generations were also included from patient interviews. So this is the first one -- the DASH did not really have patient interviews when they generated the items whereas this one did.

It was validated via a prospective study done with distal radius fractures and scaphoid fractures in 101 patients. There's two sections, for a total of 15 questions, very manageable, 15 to 30, manageable. Measurement concept, wrist function; consists of two domains, pain and function. Functional scoring is out of 50. The pain scoring is out of 50 as well. And it's average of both -- average -- the two sections are then averaged in a normalized scale of 1 to 100 just like the DASH. Less is better. And MCID, again, has been published and to be about 12 for the PRWE.

So I started to use the PRWE to supplement my DASH in order to really focus on the wrist. The positives? It's developed with patient interviews; it's region-specific. But there is -- it's short and quick. The drawback? It's validated from fewer centers than the DASH. So it's not, certainly, it's not as prevalently accepted as the DASH is, and the advantage of

it being region-specific is also a disadvantage. That means you can't really use the PRWE by itself technically because it doesn't talk about general, overall health. And you can see the two domains are about pain and function. It really misses the third domain of quality of life. Again, no hand dominance as well.

The other hand survey that's been developed out there that I looked at very carefully to employ is out of the University of Michigan. It was developed by surgeons of the University of Michigan Hand Center, called it the Michigan Hand Questionnaire. And it's developed through existing -- so their item generation for the questions were through existing questionnaires out there and where the investigators pooled questions out of existing questionnaires that are hand-related. Any questions pertaining to the hand were then incorporated into the MHQ, and then a hand patient panel developed additional items.

And it was evaluated through patient, surgeon, and therapist panels to categorize the scales. It was also given to psychometricians to identify unclear and redundant items. I have to admit, you know, when I read that word, I had to look it up in a dictionary. Wasn't something I learned as an orthopedic resident. Factor analysis was used to pare down the questionnaire. And the validation of the survey was done by the same group of researchers.

This MHQ has six sections, for a total of 65 questions. The

overall concept is evaluation of the hand, not the upper extremity, but just the hand. Multiple domains: function, activities of daily living, pain, work performance, aesthetics, and patient satisfaction. The actual scoring, this is the algorithm. You can just take a look at this algorithm, and it's not meant to be read, but it is absolutely confusing. It's very confusing to me. The MCID has been done by the same group who developed this. And the MCID actually has different -- there's a MCID for every category. So there's six different categories; there's six different MCIDs.

So when I looked at this, I thought the positive was that it's region-specific, it's very detailed, there is hand dominance involved; this is the first one that had hand dominance in it. The drawback? It doesn't add any new clinical assessment. There's a lot of psychometric assessment, which is good. The scoring system was very confusing. And it's really time-consuming. Patients did not want to fill this out. Sixty-five questions, you know, is way too burdensome for my patients.

The other thing I noticed as I was going through this was here's a problem. How about modern activities? None of these surveys looked at this. Typing, mouse, cell phones, digital cameras, taking your money out of a wallet. Writing, they did; the DASH assesses writing, writing a check. And how about plugging in a USB? So some of these very common things that we routinely do now actually is not in any of the surveys that we talked about just now.

So as one of the reasons for me to see if there's a difference in the functional outcome versus radiographic outcome, I wanted to include some of this stuff. So we included some of these questionnaires into the survey of our patients. Some had crooked wrists and some had completely straight wrists, but we wanted to ask them how functionally different they were.

And these are the different -- some of the things that were asked on this MASS07. Later on, after we developed this, we actually did validate it. So it was intended for clinical research assessment, specifically for the wrist, and to assess more modern activities than any of the other ones we talked about. Surgeons were questioned for item generation, and it was pilot tested with patients after initial development. Intention was to produce a short survey to evaluate wrist and hand function. It's 10 questions validated through 42 volunteer patients. And the scoring is from 1 to 10 for each of 10 questions. And, again, it mimics the PRWE as well as the DASH, going from 0 to 100. MCID, no one's done the MCID yet.

So the positive is that it addresses activities that impact quality of life in the modern age, short and quick; scoring method is straightforward. Drawback is it may not be appropriate for all patient populations, and older patients may not use a cell phone or handheld devices. And, again, no hand dominance. We did use this in one of our publications on whether or not wrist motion actually affects function. And this was also how it was validated.

Now, at this point, this is where I was awarded this Bunnell Traveling Fellowship that allowed me to sort of pursue another idea of higher learning, principles of scholarship and develop national/international relationships.

And this is the National Quality Healthcare Initiative, which was one of my major themes that I spent throughout the year. I also looked at global innovations in wrist surgery and also developed national/ international relationships. I had a chance to travel to approximately 15 different cities in the world.

But this took me to many different institutes in our country; the Dartmouth Institute, met with Jim Weinstein; Cleveland Clinic, met with Dr. Michael Keith, who's going to be speaking -- concluding our remarks today; and in D.C., met with Janet Corrigan at the National Quality Forum; met with Carolyn Clancy at the AHRQ; and met with Faisal Mirza, and this is what he was referring to in February when I was here talking about, you know, what was next at the FDA for orthopedics. And I also had a chance to -- had a phone conversation with Patrick Conway and our regional director, Bill Kassler, for CMS -- to really get a sense of where we headed in terms of distal radius fractures.

Now, I said that I spent a week in Oxford, and I think, you know -- this is Professor Andy Carr from Oxford, who's the Chair of Orthopaedics there. And he's just the most down-to-earth guy. I met him when I was an

ABC fellow as an orthopedic surgeon, and that's why I decided to spend a week with him. Unbeknownst to me, but he receives 18 million pounds in research funding from the U.K. He's the highest funded researcher, and he's a sports medicine doc. He's a shoulder surgeon. And he was gracious enough to host my family.

But what they looked at, why PROs are important in the U.K. system is NIH covers all citizens, and they have something called NICE, the National Institute of Health and Clinical Excellence that are looking at function, and they're determining coverage based on functional outcomes. So one of the things they've deemed very important is that it's critical that the functional outcome measurements, i.e., the PROs, are developed and validated appropriately. And that's one of the reasons why they funded the Oxford Institute with that much money, not just to do that but certainly one aspect of it.

And I met with Jill Dawson. And between Jill Dawson and Andy Carr, they have developed all of these Oxford scores for the orthopedic system. The only two -- actually, the only -- I asked them why they didn't touch the spine. They said there was something good in the spine already. And they didn't touch the hand because they thought the hand was too complex. But as they developed this, they included multiple subspecialties, they had a joint collaboration with the Department of Public Health as part of this; again, sort of the same idea as the DASH, when you have a large institute

collaboration. And the purpose was to create these things that were patient-centered and specific.

So why did they develop their own? And I asked them this question. Really, they found key discrepancies in the PRMs that they -- PROs that they had on the table. The data depended on the surgeon's judgment, which could lead to bias. They didn't have shorter, more specific simpler tools, and that patient involvement -- and this is one the key sort of pet peeves on their part -- patient involvement was not there at the get-go for a lot of the questionnaires that we're using. And so they wanted stuff that was specific, that had a large burden on patients, patient involvement, again, scale of development with a lot of stakeholders to develop these things together. And these are some of the key validation components that I think most of the PROs actually have good validation components.

So, in summary, here are some of the comparisons: The DASH and the quick DASH, PRWE, MHQ, and MASS07, some of the MCIDs. And the number of questions -- I mean, really, the key one is that the MHQ really has 65 questions, and it really is burdensome for the patients. The validations were pretty similar. The only additional thing is the MHQ was developed with a psychometric content validity that none of the other ones do. The negatives I think we talked about already. Many of these did not have patient involvement in the item generation.

Thank you.

(Applause.)

DR. MIRZA: And thank you very much. And I did want to mention that I was a little selfish in letting Charles go on a little bit longer, and that was because I did like the fact that he was bringing out the issue of patient involvement with patient-reported outcomes and the fact that the validation of PROs is just as important as understanding the target population and the MCID or that minimal detectable change that's important. And so bear with me.

Now we have Dr. Richard Coutts, who will discuss "Arthroplasty Outcome Instruments" from the University of California, San Diego. Welcome.

DR. COUTTS: Okay. Well, thank you. It's a pleasure to be here, and I'd like to thank the organizers for the opportunity to update my own information about this whole subject of PROs. It was an educational effort to put this talk together on my part.

I have no conflicts to report with regards to this particular subject.

And I thought I might start out by pointing out that 50 years ago, when I first went into medicine, the rule was that you didn't evaluate patients' outcomes by asking them how they were doing because it was considered unreliable. We've come a long way since then, because the reality is that patients can report on what they do and how they feel, and it was

simply a matter of figuring out how to quantify that and validate it. And I think that's been the history over these 50 years.

Now, I've been asked to talk about arthroplasty outcomes instruments, and I thought it would be important to point out that total joint arthroplasty of the hip and knee -- and I'm not going to discuss the shoulders; it would make it too big a subject -- it's the most frequently performed surgical -- elective surgical procedure in the world. In the United States alone, in 2009, there were almost a million total hips and total knees performed. And the numbers keep going up. And it's projected that by the year 2030, there'll be a 40 times increase in these numbers. This is mostly due to the increase in the elderly population, but also the increased obesity rate that we're experiencing in this country. So maybe we could lower those numbers if we could get a handle on obesity, but that doesn't seem to be the trend.

And I think it's important to point out that not all patients are satisfied with their operations. The vast majority, yes, but not everybody gets the desired outcome; more true in the total knees than in the total hip population. But that's why it's important to be able to document this and to measure it.

The principal reasons that somebody undergoes a joint replacement is for pain relief -- number one, pain relief -- but also for improvement in physical function and to improve their quality of life. And so those are the central domains that we would want to cover in an outcomes

instruments.

So, therefore, the attributes of a patient-reported outcome for arthroplasty patients would be to measure pain, stiffness, physical function, health-related quality of life, impact on mental and emotional well-being, activity, and participation. The difference between activity and participation is activity is being able to do something whereas participation is being involved in something along community lines, being out in the community. And I think you'll find that there's going to be a fair amount of redundancy in many of these talks because the subject matter is very similar.

So the recommendations that have been proposed for patient-reported outcomes in arthroplasty is that, of course, it should meet the minimum standards for validity, reliability, and responsiveness. It should include a summary of measures of overall physical and mental health that are key predictors for patient-reported outcomes, provide data that is useful and practical for the clinician decision making, demonstrate high levels of responsiveness, minimize questionnaire length so as to maximize response and compliance, and of course, be comparable with other existing instruments.

So again, validity, reliability, responsiveness -- others have said this -- measure what it's supposed to measure, give the same results in test and retest, and detect meaningful change.

But as other speakers have mentioned, a lot of things can

influence what the scores are that people will get on these outcomes instruments. Your score will vary depending upon your age; older people score differently than young people. It matters whether or not you're scoring for a group as opposed to an individual. And patient/physician relationship -- we're obviously interested in the individual, but from an industry perspective, we're interested in how the group does as a whole. I think Laura Tosi more than adequately described how sex can affect the responses to these outcomes instruments. Time from the intervention also influences how they score as well as what type of intervention was undertaken. And then the whole issue of comorbidities, I'll speak to that in a little bit greater detail.

That really is related to the whole issue of risk adjustment. It's been shown that someone with advanced age and two or more comorbidities will have a negative correlation. These are four hip scores that were compared, and the patients who had two or more comorbidities are out here. This is no comorbidities. And you can see the impact of comorbidities that they have on the score.

Classically, in arthroplasty, wrist adjustment has been done with the Charnley classification. John Charnley developed this very early in his career, and it was essentially three categories, but the second category has now been modified to break it up into two. So an A patient is your healthiest hip arthroplasty patient, and they just have one hip that has osteoarthritis. B classification is when there's bilateral disease. And we're

just talking about hips here. But now they have determined that there's a B-2, which is an individual who has a previous total hip on one side with the osteoarthritis on the non-treated side. And then Class C is where multiple joints are affected or if there are chronic diseases that impact ambulation, specifically. And that has shown to correlate fairly well with what the expected outcomes can be.

This was a study performed by Garellick in which they looked at the Harris hip scores as well as a Nottingham Health Profile, which is a generic form, and it highly correlated with Charnley classification. It's probably hard to read, but that says 100 and then 89 and 86 as you go down in the Charnley classification.

There are basically three types of patient-reported outcomes that are used for arthroplasty patients. There are the generic ones -- and I'm not listing all of them here -- there are over 30 outcomes instruments that have been used in studying arthroplasty. The generic ones cover the general health and mental health of the individual. And the three that have been used most commonly are the EuroQOL 5D, the Short Form 12, and the Short Form 36. And I think most of you are familiar with these different outcomes instruments.

Then there are intervention-specific outcomes instruments, the Harris hip score, the Oxford hip score, and the Oxford knee score, and the Knee Society score. And there are others, but these are the ones that have

either been used historically the most or perhaps maybe the best to use.

Then there are the disease-specific instruments, the WOMAC and then the HOOS and the KOOS, which you've already heard of, and they also have the physical function components, which are shorter forms. And the HOOS and the KOOS are a WOMAC plus a whole bunch of other questions. But it increases the question number significantly, up into the 40s.

The generic instruments are predictors of patient-reported outcomes, and they discriminate levels of general health and comorbidities, and they are helpful in the respect of being able to predict the outcome that a patient will have if you use it in that fashion before they have their surgery.

The Swedish Hip Arthroplasty has been collecting patient-reported outcomes data since 2001, and they've determined as a result of their studies that the Charnley category, the gender and the anxiety and depression scores are all highly correlated with what the patients' outcomes will be.

Disease-specific instruments are much better at discriminating orthopedic disorders or changes in arthroplasty patients. The generic forms don't give high degrees of discriminatory information with reference to the specific joint that you're looking at or the procedure.

I've been involved with the California Joint Replacement Registry. We're attempting to develop a Level 3 joint registry in California, which means that we're going to be asking patients to fill out patient-

reported outcomes forms. So we did a fairly extensive review of outcomes forms available. And you can see it's summarized here in this form. There are the three generic instruments that we thought we might use and then the disease-specific instruments. And these are the number of items in these different questionnaires, and you can see they vary from a low of 6 to a high of 42.

And our assessments are listed on the right. All of the generic instruments were probably useful, but the SF-12 and the SF-36 are felt to be more valid, responsive, and efficient. The WOMAC is valid, responsive, and widely used, but it has a fair number of questions, and there's another reason not to use it, as I'll mention in a minute. We actually felt that the Oxford hip and knee scores were probably ideal for our purposes.

As I mentioned, the SF-36 and 12 are probably the most studied and the most valid and reliable and responsive of all of the generic forms. The SF-12, it works great with large groups but doesn't work so well with a small sample size.

The Harris hip score and the Knee Society score have been used historically most commonly probably because they were one of the first, and so they've been around a long time. But in more recent years, I think the WOMAC and the SF-36 have slowly but surely been taking over.

It wasn't until recently that the Harris hip score was actually validated, but other studies have shown that there are significant ceiling

effects. And we haven't mentioned that particular aspect of these forms. You can have ceiling and floor effects, where you have large numbers of patients who get the top score or the bottom score, and there are probably differences between those individuals, but the forms are not sensitive enough to pick up those differences.

Wamper looked at this particular effect with the Harris hip score, and showed that 31 out of 54 studies they looked at, the Harris hip score had significant ceiling effects.

And then there's the issue of the burden of the questionnaires. Short questionnaires will improve compliance, response rate. Singh looked at this and found that if you only had 12 items, the patients didn't feel particularly burdened. If you had 24 items, it was the upper comfort level. Greater than 25 items, there was less tolerance. And 50 items, you start getting attrition and missed questions. And over 80 items, then you have patient fatigue.

And you can see this in this review where they showed that the SF-12 had a higher net response rate of 75% compared to the SF-36, with 63%, and the amount of time that it took to complete, 7 minutes for the SF-12 and 14 for the SF-36. The same is true for the Oxford knee versus the WOMAC.

The question is how robust are these various outcomes instruments for arthroplasty evaluation? This is a report of a meta-analysis

performed by Alviar, and they looked at 28 outcomes instruments, and you can read all of the different aspects of these instruments that they evaluated.

No instrument satisfied both factor analysis and Cronbach's A. Cronbach's A is a measure of the internal consistency of the instruments; 24 out of 28 had indeterminate ratings for responsiveness. And the only responsive instruments that they could identify were the WOMAC pain and function and the SF-36 physical function, general health, vitality, and mental health. So even here, they had to get down into the different domains in order to figure out whether they were any good.

Total joint replacement, or arthroplasty, has a very large effect size, and so we really don't need a great deal of ability to differentiate small differences when we're looking at these patients. We know that they pretty much do well. But there are circumstances where you'd like to drill down and find out differences, and I think that's what the FDA is probably currently interested in and why we're having this conference about the important differences.

If you wanted to compare a surgical approach, for instance, a minimally invasive versus a standard approach, the Harris hip score isn't going to be able to tell that to you. Or if you wanted to look at specific surgeries like whether you resurface the patella or whether you retain the cruciate or sacrifice it in a knee replacement, the Knee Society score is not going to be able to differentiate that. So we need these clinically important differences

to determine the power of our studies and to be able to differentiate these differences in approaches.

But as other speakers have pointed out, the MCIDs have a great deal of variability; it depends upon what the anchor is that you use to develop it, patient expectations can influence it, and depending upon the intervention, it can be a different score, for instance, if it's a surgical or a medical condition that you're evaluating. The MCID has been determined for the WOMAC in total knee replacements to be 23 for pain and 20 for function. I'm not sure what these units are referring to, but in another study of osteoarthritis and the use of NSAIDs, the MCID was 20 for pain versus 23 and 9 for function versus total knee replacement.

MCIDs have been determined only for the WOMAC and the SF-36 in these arthroplasty instruments, and you can see the numbers there. But I don't know how reliable these are because it would depend upon exactly what circumstance you're measuring.

And then there's the issue of accessibility. The WOMAC and the SF-36 and 12 are not free. Royalties have to be paid for these particular instruments. And I think that reduces their usefulness. And we would hope that any outcomes instruments would be in the public domain and available to everybody.

There are new instruments coming down the road. The Knee Society has just developed a new outcomes instrument, and they've published

that. The Hip Society is working on a new score. And we've heard earlier about the PROMIS effort that the NIH has been supporting. It has not been evaluated for arthroplasties or even for musculoskeletal conditions, to my knowledge, but it has promise in that area.

So at this point, in terms of doing arthroplasty evaluations, the current recommendation is that you use a combination of a disease-specific instrument as well as a generic instrument, which would cover all of your bases, but try to keep the number of questions low using instruments that are validated and responsive.

Thank you.

(Applause.)

DR. JONES: Now open for questions.

DR. DAVIS: It's not a question as much as just a clarification for people. So the issue with the WOMAC is that version 3.0 Likert version is available in the public domain. So you can't -- Nick Bellamy will not give it to you anymore, but people have copies, and as long as you specify you're using 3.0, it's fine. It's version 4.0 that he copyrighted and makes you pay for. And we've gone through this because we had to deal with it with the OARSI-OMERACT things, so you just have to be very careful that you're specifying you're using 3.0.

DR. JONES: So one of the questions that comes to mind for me, and it was brought up this morning and then again this afternoon --

DR. MIRZA: He has a comment.

DR. JONES: Yes, sorry, he's going to comment to that.

DR. HOCHBERG: So let me make a brief comment. So I'll declare a relationship here, which is that Nick Bellamy and I are very good friends. And if you're an academic investigator, Nick will provide you with the copyrighted, trademark version for academic research. So royalties are charged for companies, which presumably have much deeper pockets than academic investigators for the purposes of collecting data for licensing products and for submission of regulatory dossiers.

DR. JONES: Okay. To get back to my question, because, Dick, this goes to part of what you were saying. Years ago, I asked Bill Harris a question. I said, you know, how do you handle when someone has bilateral disease. And he says, "I don't include them in my research studies." And that was how he got around it. And I'm hearing it from this morning and this afternoon that, you know, it's particularly of issue for patient-reported outcomes. A patient that has one -- you know, both knees affected are miserable even after their total hip replacement if their other contralateral side is hurting. Can you comment on that?

DR. COUTTS: I think that's what makes this such a difficult, somewhat confusing area to evaluate. When you have a patient fill out a Harris hip score or knee score, it's intended that they will give responses relative to the joint under question. But, of course, they don't have just one

joint. God was very good to the orthopedic surgeon, and he gave everybody two joints of almost everything.

(Laughter.)

DR. COUTTS: And lots of joints. So if you're asking somebody about their right hip, but they also happen to have arthritis in a knee or both knees or the other hip, or they've got a bad back, that's going to influence their answers to those questions. And so it's very hard to tease out the -- how you're doing with that specific joint that you're really interested in because you just operated on it, for instance, and you want to know what the result of your surgery was. But it has to be evaluated within the context of the whole person. And so I think we just have to be cognizant of the fact that we are dealing with whole people and all of the other foibles that relate to them. And that, in the end, is probably the most important evaluation is just exactly how is that person doing. But then you have to know why they might not score as high as maybe somebody else who only has one joint involved and you've taken care of that and you've restored them back to almost normalcy.

DR. HOCHBERG: So actually, there are now studies which have done knee-specific, at least WOMACs, ICOAPs, and KOOSs, so the Arthroplasty Initiative, which is an NIH-funded longitudinal observational study, has data over 8 years on knee-specific instruments. So because we look at individual knees based on radiographic outcomes, MRI outcomes, also clinical outcomes using -- asking people to complete the instruments for their right knee

separately from the left knee.

AUDIENCE MEMBER: Philosophical question somewhat at Charles, but I think from all of you. And that is I trained with Marc Swiontkowski, who is a generalist and prefers a overall outcome measure of musculoskeletal health. I did my fellowship with Sandy Kirkley, who developed three independent shoulder outcome measures. And they were philosophically opposed, obviously. And so I got the sense from you, Charles, that you thought if you have a joint-specific measure, that that's preferable to a general one, and use the general ones for something where you have a lower, you know, burden of disease in terms of, you know, supracondylar elbow fractures or something in an adult, you know? What's your preference?

DR. DAY: That's a good question. No, actually, that was just my hypothesis, so my hypothesis with that study was that if I looked at something more specific, then I might find a difference between a radiographic alignment and functional outcome. And it turns out, no, it didn't. So even though I used a more specific one and I used others, and I also included a generic one, there was no difference. So I'm still not -- so at this point, I'm not sure what the utility is. So it hasn't -- at least in my own studies, they have not demonstrated a difference between the more specific ones and the DASH, the more generic one.

MS. STONE: We have another question from the webcast. This

is from, I'm sorry, I'm going to mispronounce this, but Roxolana Horbowyj, medical officer here in CDRH. How is MCID or MID being differentiated from the traditional delta, defined as the clinical significant difference between comparators that have been used in device clinical trial design and decision making? Delta can be different depending on comparators and comparison, for example, superiority, non-inferiority, et cetera.

DR. COUTTS: Wrong panel.

(Laughter.)

DR. HOCHBERG: I'll take a crack. So the delta would be just the -- let's say the absolute difference between the treatment groups. So in a superiority trial, you would test to see whether that absolute difference was significantly different from 0. I'll pass to the biostatisticians in the group because I'm not a biostatistician, but you could either do that with, you know, traditional testing; you could do that by looking to see whether the 95% confidence interval nominally included -- the nominal confidence interval included 0. And that would give you your delta.

Now, how do you set the delta for a non-inferiority trial? You do that after discussions with the Agency, because at least in the world of drugs, the landscape, I think, is littered with trials where there was not an adequate discussion beforehand, the sponsors came in for approval to try and get non-inferiority labeling, and there was then this sort of post-study discussion that, well, you know, your lower bound of the non-inferiority was

not really close enough to 0.

So if you look at the NSAID literature, most of the studies will choose 10 mm on a 100 mm scale. But you know, more recently, at least in my experience as a consultant, it's been suggested that that should be 8. I don't know what it is for devices.

So these are unrelated to what the minimal clinically important difference. And I think it's important for the Agency and for sponsors to recognize that the MCID is a individual -- is something which is measured individual for the purposes of determining the proportion of subjects enrolled in a study who meet the MCID in order to be able to compare that between treatment groups and not to interpret the delta as to whether it exceeds or fails to exceed the MCID in making decisions from a regulatory standpoint. And I think that is consistent with what was said this morning by some of the other speakers.

DR. KEITH: Hi, this is Mike Keith. Got a question for any and all. I have a lot of trouble doing clinical trials especially using outcomes instruments because so much of it is pain score-dependent. And in my hospital, every patient comes in with six medications and is in a pain management program, sent there by the people who started treating them. And I'm supposed to collect instrument outcomes information on these folks while they're still being managed.

From a practical point of view, how is FDA going to give

guidance to this conflict of, you know, multiple management, multiple drugs at the time when devices are supposed to show pain reduction?

DR. JONES: Do you want --

DR. COUTTS: Well, I can just relate that I've been involved in studies where patients are required to go through a washout period beforehand and then you get your data at that point in time. That's sometimes viewed as being cruel and inhuman punishment because you're really requiring the patients to go through a period of discomfort. But from a scientifically valid perspective, that's the only way you can really deal with it.

DR. HOCHBERG: The other way -- oh, go ahead.

DR. DAY: I was just going to say, yeah, the only way -- I mean, it sounds like you can't get away with it, but the only way we can do that is to exclude, so we exclude patients that are on multiple pain meds right off the bat, but it sounds like in your hospital that'd be difficult.

DR. HOCHBERG: So the other way you could do this, I guess, and this would be in the context of the fact that you are doing a randomized trial, okay, is that you can either randomize people and hope that randomization will give you comparable groups with regard to the background medications that they're using and test the efficacy of your device as an adjunct to the background therapy.

But if you're concerned about specific types of background therapy, let's say opioid analgesics, then you can follow the suggestions of

one of the speakers this morning and stratify on those who are using opioids versus those who aren't using opioids and then randomize within the strata. So that will help you in terms of a randomized trial of a specific device.

But if you're just looking to see in an observational study whether your device is helpful for pain relief, you know, what's done -- I mean what could be done is that once your device is, let's say, implanted or inserted or used in the treatment of the patient is when they're brought back at intervals for follow-up, they can be washed out of specific medications if they're still taking them at the time of the follow-up visit.

So this is what's done in structure-modifying studies, for instance, in osteoarthritis, where the Agency wants to see whether there's a benefit on pain, because at the moment, you can't get something approved for the indication of structure modification unless you also have an improvement in pain. There may be disagreement about that, but that's the reality.

And what you do is basically discontinue pain medications five half-lives prior to the visit on which you're going to measure pain at an outcome visit, so it would be a 3-month or a 6-month or a 12-month visit; you might discontinue the analgesic medications, you know, anywhere from 2 to 7 days prior to that visit and then measure pain to see whether the compound that's being tested for effects on structure would have an effect on pain as well. It's not optimal.

DR. JONES: So this morning we heard the statement that some people were cheating as far as sample size determination. I want to bring that to something that the three of you spoke about, and that was validation. One of the things that we learned this morning, which many of us are aware of, is sex-based difference; there are cultural-based differences. The validations are frequently for the physician instruments, not so much for the patient-generated ones, and not validated necessarily for preop as well as postop. And so when we call something validated, are we saying that they're validated along all types of validation, or are we just picking and choosing one and saying, oh, it's validated. And all three of you spoke about that, so if you could comment?

DR. DAY: So I can only speak to the one that I had validated, the MASS07. And now that I've had some experience in Oxford and looking at some of the things that should go into a patient -- you know, I wish I had had that experience before looking into something like this.

But, essentially, the validation process, I thought it was fairly simple. I mean, I just -- I had developed the 10 questions I wanted to ask, my questions that was not being answered in any other outcome measures. And then I just looked up how the DASH was validated, you know, and simulated a -- especially the change in sensitivity, the responsiveness to the limitation -- we simulated a limitation by putting somebody in a soft splint versus a hard splint, and then tested their functional outcome of their hand. And so that's a

very concrete simulated condition, and patients were randomized to whoever got the hard splint first or the soft splint, and what have you. And so there was a distinct correlation between range of motion, between the, you know, no splint, hard splint, soft splint, and also the functional outcomes, and then also the validation.

So for my particular one that we validated, we did all three of those, and that was what was necessary to get it published.

DR. MIRZA: Sorry, if I could just make a quick comment about validation, I think that's very important. And I just wanted to have Heather -- she's our epidemiologist who's a fellow here who is working on the research project that we're looking into. And one of the first steps of the project is to identify the validated PROs. And so I'll just ask her to address what steps we're doing in terms of identifying the validated PROs.

MS. STONE: Hi, thank you. So what we have found from our work with Art Sedrakyan and others is that it's a rather complicated process in the sense of doing a very extensive validation. To just say that something has content validity is quite simple; you know, does the question ask what it's intended to address. But when we get into the issues of construct validity, internal consistency, reliability, and all of the latter psychometric properties, it becomes more difficult.

Now, many of them individually are, you're correct, not that hard to actually identify. There are simple statistical tests that can be done,

and if those are reported in the literature, then you can. And what we're trying to do is basically to comprehensively assess what is the evidence for each instrument in terms of this wide range of psychometric properties and validation. But not a lot has been published on it, and it's very scattered. So to try to assess not just is there this inherent sort of content validity, but what is the construct validity, how was it assessed in terms of comparable instruments. And then I think we're really missing the component of how does that change looking across different variables. That information, as was sort of presented earlier this morning, is not very much in the literature at this point.

DR. MIRZA: Great. Thanks.

DR. DAY: I would just say that when I was looking up at least the upper extremity ones, I mean, the validation seemed to be fairly uniform. The problem with the upper extremity instruments were primarily in the question generation, that a lot of the questions were not generated with patients in mind. They were all just like the one I developed; that was all my idea. I thought that -- it was a hypothesis. I was testing. And it turned out to not be effective and didn't really get me the answers I wanted anyways. But it was sort of -- I think the item generation, getting patients involved is really -- I think it's a really critical part of PROs.

DR. MIRZA: If I could just ask, one thing that was brought up when we were discussing the project, and I wanted to bring it up now because

we're getting into sort of the clinical aspects of anchoring, the statistical -- because we just had our statistical panel; now we have our clinical panel. One question is connecting the dots between the .5 MID and the MCIDs that we're talking about, because that question was raised by a couple of clinicians as to why is -- what is .5 MID. And now we're talking about an MID of 10, 12, 9. And I think it's just based on one is a subdomain and in the other situation you're measuring the total score. And then the other question is should we anchor a PRO to an actual functional clinical exam as part of that validation? And this is not something that we can do now, obviously. Historically, we already have all this published data. Is that something we should be doing? And I know we've talked about that, Charles and Mike, with Dr. Swiontkowski as well, so --

DR. HOCHBERG: So I would say that the .5 represents the clinically significant change, not the minimal clinically important difference, but what was considered by the presenter, if I understood correctly, to be the clinically significant, which would parallel, let's say, the statistical significance. So statistical significance is related to sample size. You can get, you know, a large enough sample is the data we're shown this morning, and you can show that a 2 mm difference is statistically significant, but that's not clinically important, I would guess, unless your scale goes from only 0 to 4 mm. But if your scale goes from 0 to 100 mm, a 2 mm difference is not going to be clinically important. So what the half standard deviation is, is a clinically

significant difference, okay?

But that's different from the minimal clinically important difference, which is, I think, the subject-anchored difference in the measurement scale which corresponds to what the subjects feel is a minimally clinically important difference. So are they improved, yes/no; is that improvement slight, you know, minimal improvement, moderate improvement, or marked improvement.

So what is the delta for the people who have minimal improvement, okay? You can talk about minimal perceptible clinical improvement. You can have a minimal clinically important improvement or a minimal clinically important difference. But none of those are going to reach the half standard deviation. Half standard deviation is going to be much larger on whatever the measurement scale is, I would guess.

DR. COUTTS: Maybe like many of you, I've been confused by some of the numbers that have been thrown out here, and so I had a sidebar conversation with Dr. Sloan after the morning session trying to figure out what was the relationship between the numbers that I had seen in the literature of an MCID of 20 or 30 and this .5 standard of being above or below to signify a clinically important difference.

And if I understood it correctly, the only difference is dividing that larger score by the standard deviation of the group. In other words, you take the difference in the score -- and the example that was given earlier in

the morning of a 7-point score. And let's say that the patient had an intervention and there was a difference of 3 in one domain and then you added that all up and got the difference, and so you would end up with a 20 or 30 total. And then what they did was divide that by the standard deviation of the group. And that's going to give you a number between 0 and 1.

That's my understanding. I don't know if it's correct. If somebody -- if one of the statisticians out there can further elaborate, I would appreciate it.

MR. BOST: (Off microphone.) It's --

DR. COUTTS: Means?

MR. BOST: It's the difference between means.

UNIDENTIFIED SPEAKER: You need a mike.

MR. BOST: It's the difference between means divided by the pooled standard deviation.

DR. COUTTS: Okay. So it's the means, not the number difference, okay.

DR. DAY: I would just comment on that last point you made on whether or not we need something sort of -- something observational to look at how functional outcome is. I mean, in my -- again, as part of that study, trying to figure out the relationship between radiographic outcomes and functional outcomes in addition to looking for all these things, I actually looked for something that was objective as well.

And the idea is, you know, like -- and I tell this to my research students all the time -- is I may feel -- if somebody asks me if I -- could I still play basketball, I'd say yes. But I'm thinking I'm playing basketball when I was 18, 30 pounds lighter. And on the court now, I may feel like I can -- if you ask me a question can I play, I'd say I can play, but I'm not sure that that's real.

And so I found something called a Jefferson Taylor, which was developed in 1969. It's just something that's a whole sequence -- it cost 300 bucks, you put a whole bunch of beans into a can, and you take it out -- put paperclips into -- but you're timing the patient doing these things while, you know, while you're actually observing them and timing, so it's a timed hand performance test. So I added that in addition to the other PROs.

DR. MIRZA: Okay. And I think what you're referring to, your ability to play basketball may be an intuitive heuristic bias, according to Daniel Kahneman.

But I think -- were there a couple of comments about -- okay, do you have -- yeah, go ahead, sorry.

MS. SPEERING: Yes, I have a question. I'm Leann Speering from Wright Medical Technology. And the general health outcomes are widely cited in the medical literature. Researchers often want to collect them to be relevant and comparable to other studies. So to play the devil's advocate, do the general health outcomes measures perpetuate themselves? Do they actually add utility to the field? Or do we just collect them and report them

because everyone else did and we need to have some of that data in our study as well?

DR. COUTTS: I can answer with reference to the arthroplasty instruments. I think it's important to get these generic, general health questionnaires filled out because one of the things they capture are items like depression. And studies have been done now showing that patients who score high on their depression score do worse after joint replacement. They tend to have persistent pain. So it is important to get these general questions which capture more or less the total body position of the patients as opposed to the joint-specific or disease-specific instruments, which tend to focus mostly on the joints themselves.

DR. HOCHBERG: If you want to know the cost effectiveness of your intervention, then you should include a general health outcome measure, which will allow you to calculate -- let's say will allow you to estimate utilities and allow you, therefore, to look at cost effectiveness and cost per quality-adjusted disability-year or life-year. So if you're operating in certain health systems, not the United States at the moment but others, and you want your intervention once it's approved by the regulatory authority to get paid for, you have to demonstrate that it's cost effective at a certain level.

And in order to do that, you're going to need a general health outcome measure as part of your studies.

DR. JONES: Okay. Thank you very much. We'll adjourn this

session. Thank you.

(Applause.)

DR. MIRZA: Thank you, everyone.

Now, we are scheduled to have a break, but I had a suggestion that if we move into the Industry Perspective, since we're running about 10 minutes early, we can actually finish early. So if no one objects, then why don't we move into the next session?

And -- I'm sorry -- we also for tomorrow's session, we have breakout groups based on different joints. Like, there's a hip session, a knee, shoulder, elbow, wrist, hand, foot, ankle and pediatric disability breakout group. And there are sign-in sheets that will be -- before you leave today, please make sure you sign in as to where you intend or plan to go. And we have different room assignments, each of which are being webcast as well, so if you can remember to sign in.

But, otherwise, I think I'd like to go ahead and proceed with moving on to the Industry Perspective, and we'll get started in a couple minutes once we make sure we have all the slides up.

(Pause.)

DR. MIRZA: -- begin the session. And I'll ask Dr. Charles Turkelson to come up. He will be moderating the session.

DR. WYRWICH: Hi, I'm Kathy Wyrwich from United Biosource Corporation, and I'm here to --

UNIDENTIFIED SPEAKER: (Off microphone.) You need to move the microphone.

DR. MIRZA: There's two of them. This is louder than that one.

DR. WYRWICH: This one's louder?

DR. MIRZA: Yeah.

DR. WYRWICH: So if I use this one, it's better, and perhaps I'll use this one.

Okay. So I'm here to give an overview of the industry perspective in terms of PROs and particularly in terms of interpretation. And it's from the interpretation that we get to the whole MID term.

I need to always present this slide because my company, United Biosource Corporation, is a wholly owned subsidiary of Express Scripts, but need to make sure that everyone's aware that we are firewalled and, in fact, our data is protected from anything that Express Scripts learns.

The outline of the things that I want to look at are the FDA approaches that have been clearly delineated in the guidance for PROs for interpretation of patient-reported outcomes for labeling and promotional claims. And although labeling and promotional claims happen way down the line in the process, it's always very important to keep the end in sight. I'm going to talk about the idea of a responder definition, which we've heard talked about in several different ways throughout the day; anchor-based methods for determining that responder definition; the role of distribution-

based methods that Jeff Sloan talked a little bit about; what happened to the idea of the MID, how that has disappeared from regulatory guidance; and the role of cumulative distribution functions.

So, first, in terms of an introduction, we're going to be looking at pre-specified patient-reported outcomes, and we've already this afternoon seen many of those described. And it's important that all PROs be pre-specified in that we find out about what happened in terms of change over time in the PROs, whether it's specified or not. That information on the PROs really is useful and helps to give a total picture of what has happened. In response to the question earlier, are we're just measuring this for the fun of it, the PROs really give us insight on what the patients report.

Interpretation is where this whole idea of the MCID came from. Gord gave an outline of one of the first studies that he did, the Jaeschke study in 1989. And what that group in McMasters realized was that with very few patients, they could achieve p less than .05. But they had no way of interpreting what the PRO scores or the PRO instruments they were using, whether that was important to the patient. So interpretation is what we're all searching for with this MID concept, which is, in fact, is a term that is not used anymore.

Interpretation of PRO endpoints follows very similar considerations for what happens with other endpoints and is used to evaluate the treatment benefit of a product.

So here's the definition of interpretation that's from the Consort Group. It's the extent to which the results and findings are consistent with the original hypothesis, including considerations for the robustness of the results, for drawing conclusions and making recommendations, what do we learn after we measure the PRO across treatment groups.

Now, for many years, I was an academic at Saint Louis University in the Department of Research Methodology and Health Services Research. And in 2008, I left academics and joined industry at United Biosource Corporation. And for us, the rules of the road -- which we try to give as much input when guidances are being developed and formulated and when they're in the draft version -- the rules of the road is the PRO Guidance. And, in fact, this guidance, as we see at the very bottom of the page, was released, the Final Guidance, in 2009, and was originally developed and released in Draft Form in February of 2006, and it applies to CDER, CBER, and CDRH.

And what does the Guidance tell us? I think this thing works better for men than it does for women. Okay.

UNIDENTIFIED SPEAKER: (Off microphone.) Sex differences.

DR. WYRWICH: Yeah.

In the PRO Guidance on page 7, we get a picture that's referred to in the industry as the wheels and spokes on how PROs are developed. We've heard briefly earlier today about content validity. And that begins at

the beginning by hypothesizing the concept we want to measure and then adjusting the conceptual framework and drafting the actual instrument, which definitely should include patients, confirming that concept. And then over there in the red square that I've added around step 4 is where we look at collecting, analyzing, and then interpreting that data.

And in that process of interpreting is when we develop responder definitions, test those responder definitions, decide if in the next step we need to modify the instrument, and if in modifying the instruments, we need to go back through the wheels and spokes again to make sure that we get that instrument in the best shape so that it's fit-for-purpose. And that fit-for-purpose is, in fact, label claims and promotion in my industry.

So necessary before the clinical trial data can be looked at in terms of interpretation is that we have these basic issues of validity taken care of; that there is a conceptual framework that the instrument was built around; that we have content validity that has construct validity; reliability, that it's able to detect change, which is not necessarily a separate measurement property from construct validity, but we want to actually see that happen; and that we have cultural and linguistic adaptations if we're running a trial in countries other than the country where the instrument was developed. We also want to make sure that the clinical trial design passes all of the necessary steps, we have blinding and randomization, method for handling missing data, and an improved statistical analysis plan.

So with all of that in place, then we can start to look at the issues of interpretation. The important thing I want to stress, though, is that interpretation is more than this p less than .05. We definitely want to set up our clinical trials to make sure that we achieve p less than .05, but once that happens with our PROs, how do we go to the next step and interpret that? It's very important that we do that because, often, the people that read our labels are not going to be able to understand the PRO that's incorporated. It's not a part of their experience. And so the PROs need that extra help in terms of interpretation. And this gets us to the whole idea of deciding who met certain thresholds and who didn't.

So the responder definition is the key term that the FDA would like to see in terms of interpretation. MID -- and certainly, I hope everyone understood Gordon Guyatt's comments earlier this morning that clinically is not a term to be used in this arena of looking at patient-reported outcomes because we use patients to help us interpret that, and there may be an instance where they're responding in the clinic to that, but there's not really a clinical aspect to this. So many, many years ago, the C in MCID was dropped and became the MID, and as you'll see at the end of this talk, the MID has actually been dropped also in favor of this term of a responder definition.

So a responder definition is a key to being able to interpret the PRO results. It's defined as a trial-specific important different standard or threshold applied to the individual level of analysis. It represents the

individual patient PRO score over a pre-determined time period that should be interpreted as a treatment benefit. And it's important to know that the PRO responder definition is trial-specific. We don't want to use the MID or the responder definition that's been determined for knees when we're looking at hands or we're looking at shoulders. We don't use the SF-36 responder definition to be used across all areas, but it needs to be determined within that particular trial. So if there's a particular type of patient being enrolled in the trial, we want to look in that particular patient group and find out what is the most applicable responder definition to be able to interpret those trial results.

So the responder definition is determined empirically and may vary across the target population and from other clinical trial design characteristics. So from trial to trial, we might have different responder definitions, but we always want to have empirical proof to back that up. FDA reviewers will evaluate the PRO instrument's responder definition in the context of each specific clinical trial. And regardless of whether the primary endpoint for the clinical trial is based on individual responses to treatment or to the group response, it's usually useful to display the results for the individual PRO responses.

So what that means is that you could define as your primary endpoint not change over time on a particular score but achieving a certain threshold. And so in that case, you have individual response to treatment as

your endpoint. However, it could be change over time that is your endpoint. And in that case, it really is helpful to take those group means for change over time and go back and interpret them in terms of what's happening at the individual level. And that's what the responder threshold -- responder definition gives us an opportunity to do.

There we go. Okay. So the process that we do this is through anchor-based methods. And the anchor-based methods were discussed earlier by Gordon Guyatt, that these are methods to explore the association between the targeted concept of the PRO instrument and the concept measured by the anchors.

To be useful, the anchors should be easier to interpret than the actual PRO and all of the items within the PRO. And it should be correlated with the PROs' change over time. And one of those thresholds that we often use on that correlation is the absolute value of the correlation is greater than or equal to .3. It should provide meaning or interpretation for that change, which is the whole idea of the anchor helping us to get to the concept better, and it should be intuitive.

So types of anchors that are specified in the FDA PRO Guidance are clinical measures such of $FEV_{1.}$ Has anyone here ever done a correlation between a respiratory measure PRO and FEV_1? Generally, pretty poor. If you achieve .3, that's actually really good. However, this is the example that's provided in the PRO guidance. Perhaps there are other clinical measures that

more closely approximate what's happening in the PRO for the specific area.

A second type of anchor that's recommended is a clinically reported outcome, such as a clinician global rating of change and then also patient global rating of change, which Gord talked about, but I want to extend that to the current thought from the FDA.

So on the clinical measure, an example would be the number of incontinence episodes that's been used to determine the responder definition for a PROs instrument that assesses the annoyance of incontinence. And a 50% reduction in incontinence episodes may be proposed as an anchor for defining the responder definition. And confirmation of this anchor approach in early clinical trials -- and again, in early trials, it can help provide the basis for a proposed responder definition for the confirmatory trials. So knowing *a priori* what that level is going to be in your Phase III trials is very useful.

Second type of anchor I mentioned earlier from the PRO Guidance is a clinician-reported outcome, like a clinician global rating of change. And an example I have for that is from a study looking at the quality of life, enjoyment, and satisfaction from three 8-week trials that were done using drug, not device in generalized anxiety disorder. And I would only recommend looking a clinician-reported patient measure of patient change as opposed to the patient-reported measure of patient change when there's a situation where the doctor may be a better reporter on the patient's condition overall, like a mental health condition.

I absolutely, giving where I'm standing right now, must give a disclaimer of the fact that this is a published study, but it does not represent an FDA approval for this drug in this indication.

And here's what the results look like when the clinician-reported impression of change over 12 weeks was given. And we looked at most closely what level of improvement in the mean change score happened for those who were minimally improved, and we set the responder definition for this particular instrument, the QLES-Q-SF, at 6.80 points, based on the mean of those 293 patients whose clinician reported that they had improved over time. We felt confident in these results because, although there's not incremental change, there is a certainly a strong trend, and those categories of CGI-I improvement aren't necessarily linearly separated from one another.

The third type of anchor that I mentioned -- excuse me -- is patient global ratings. And, certainly, the grandfather of all studies is one that Gordon Guyatt mentioned earlier this morning from 1989. And in that study, patients were asked a global question for each domain: dyspnea, emotional functioning, and fatigue.

So, for example, for the fatigue domain, which has four items in it, they would be asked: Has there been a change in your level of fatigue since your last visit? And patients would respond whether they were better, worse, or about the same. And those who said that they were better were given a scale like the one that's on the left. And those who said they were --

oh, sorry -- on the right. And those who said they were worse were given the scale that was on the left to decide exactly how much better or worse they were. And those who responded at the level of 2 or 3 were considered those who had had a small but important change, or the minimal level of change.

Now, another -- well, let me just go back and do some interpretation here, too. So this is the scale that they chose from. And then when people answered each of the items, like on the fatigue domain, the McMaster scales are all on scales of 1 through 7. So on the four-item fatigue scale, your score can go anywhere from 4 to 28, because there's seven response levels.

However, the McMaster folks report that out as the average per item. So they'll take the range from 4 to 28, and they'll divide it by the number of items. So actually, your scale score in fatigue for the four items in the fatigue domain can range from 1 to 7. And the .5 that Gord talked about that they found out to be the MCID, a term they used in 1989, was the average change per item. Now, an item can't actually change at .5, but it can -- two items can have a one-point change between -- across two items, one response level on one of those items can have that change. So for the 4-point fatigue domain, the MID would be a change of .5 per item or an overall score change of two points.

Another example of using a patient global anchor is a study that's very famous done by John Farrar and his group at Penn looking at pain.

And they used the numeric rating scale that goes from 0 to 10, no pain to worse pain. And they had daily entries, and this was done across 10 studies for patients with fibromyalgia and osteoarthritis. And they used an anchor that was patient-focused and had patients answer my overall status from the start of the study, and patients said whether they were minimally, much, or very much improved, no change or minimally, much or very much worse. And those who were much improved were considered to be the responders.

So then Farrar and his group did an ROC analysis and looked at where scales best differentiated how much change, best differentiated those who were much or very much improved from those who were minimally or had not had improvement. And the ROC analysis found that the change of -1.74, which we interpreted as a change score of 2, was an important change. And also because percent change has a very long tradition in the pain field, they also looked at percent change.

Now, I always want to put a warning label next to this because percent change is a very slippery slope, that someone who goes from a 9 to a 6 on their pain score has the same percent change as someone who goes from a 3 to a 2. But yet I think that the change exhibited that the effect on the overall individual going from a 9 to a 6 perhaps is much greater than what we see in going from a 3 to a 2.

So percent change is reported here and, again, has this very strong tradition in the pain field. However, always be real careful about

what's really happening when you do use that metric.

I also want to point out, because I'm going to come back to it later, that the very much improved, their percent change was close to 50%, and we see the close to 30% on the much or very much improved threshold. And we also see change scores of about a 3 for those who were very much improved versus about a 2 for those who were -- I'm sorry -- yeah, for the much or very much improved on the ROC. So these are key pain thresholds, 2 or 3, 30% or 50%.

However, Gord did do some talk in his presentation this morning about the real problematic aspects of this idea of doing retrospective assessment of change. And he talked a lot in correlations, and I'm not sure everyone caught what was happening there. But the whole idea is that when we ask patients at the end of a study about how they feel like their change has been, are they better, worse, or about the same, their responses very much underestimate their initial state and they're highly correlated with their present state. So if I'm feeling good now, I say I'm better; if I'm feeling bad now, I say I'm worse. But I don't necessarily go back and do the mental gymnastics between where I was at the beginning of the study and where I am now to come up with a real improvement score.

And one of the best ways, I think, to illustrate that is to think about the dinner you had last night compared to the dinner Monday night 2 weeks ago. And I have to ask you that, how would you rate your dinner last

night compared to November 12th, does everybody feel like they could answer that? You put it in front of patients, they'll give you an answer, okay? You know, they somehow feel that their care or their cooperation -- no matter how many consent forms you put in front of them, you put it in front of them, they'll give you answer. Can anybody realistically answer that? Well, if your birthday was November 12th or your anniversary was November 12th, perhaps you can and you can give a realistic assessment. Most of us could not come up with what we had on November 12th for dinner. We can remember if we ate in an airport last night and would probably give this a bad rating.

Patients do the same thing. And I have people raise their hand and say, well, food isn't nearly as important as pain. And I'd say, well -- there could be an argument there, but the whole idea that it's very, very difficult for patients to do this retrospective assessment of change is what I want to bring to mind, and that's what Gordon was doing a lot of figures and correlations to demonstrate earlier this morning.

So in recognition of this fact, the FDA has looked at two different types of patient global ratings. And one is called patient global rating of concept, and I've put a smiley face next to that, and the patient global rating of change, and I've put the poison face next to that just so you don't get them mixed up in any way, shape, or form.

The patient global rating of concept is a comprehensive

evaluation of a complex concept. It's non-comparative. So you're only going to rate your current condition or at least over a short time span and not have to compare back to the beginning of the trial. It's minimal to no requirements for patients to do that long period averaging.

And an example would be: How would you rate your irritable bowel syndrome symptoms overall over the past 7 days? And so patients are basically giving a 7-day cross-sectional analysis of what they have here, and we would want to get a cross-sectional analysis at minimal -- excuse me -- a cross-sectional analysis and like to see that over time, that would change about 2 points, and then understand those patients.

The global rating of change, we just talked about. That's what John Farrar and Gord looked at, and that has you doing that retrospective and trying to remember who you were at the beginning of the study.

So the rule of distribution-based methods, Jeff Sloan talked a little bit about that the one-half standard deviation. And, again, that's looking at change scores divided by the pooled baseline standard deviation, and we want to see that at about half a standard deviation; another one that chooses a standard error of measurement. But this is only to use to support the anchor-based differences you'd want to see.

And, finally, the MID is missing because it looked at the idea of group change. And it required that the group change and the confidence intervals around that group change actually be bigger than a specified

threshold.

So MID is a term that's not used at all in the PRO evaluation -- in the evaluation of PROs anymore at the FDA, but instead, we use the idea of responder definition and then also the idea -- and I'm going to move quickly through this -- of cumulative distribution functions. And this compares what happens over time.

And here's a picture from the Aricept label. And the only thing that I have added to the label as it actually stands are the three purple rectangles that show what's happened on the these three different treatments -- two treatments and one placebo group over time so that you can actually see the change score of every patient contributing to what's happened there.

And then here's a picture from the Cymbalta label. And they've actually drawn in the 50% change and the 30% change, and this is the cumulative distribution functions in reverse. So you can actually see at these important thresholds. But every patient was allowed to contribute to what happened in those pictures there.

So what I hope that we understand, or what I hope to talk about are these important issues with the FDA and what is contained in the PRO Guidance for the interpretation of change over time, and these are the topics that hopefully have a better understanding now.

(Applause.)

DR. MIRZA: Great. Thank you very much.

And now we'll have James Ryaby, Dr. James Ryaby, from Ryaby Associates give us his perspective.

DR. RYABY: Good afternoon, everyone. I guess I'm going to have the same problem. I do have a disclosure. And I think, most importantly, I also do some consulting for investment banking VC firms. And the reason why I'm just showing you this, for those of you who don't know me, is at OrthoLogic in Arizona, we started as a venture capital-based medical device company that got two devices approved through FDA and transitioned to a biopharma company, and in the middle of that, we went public. So I have experience both with the industry perspective from VC funding through public company. Then I have experience in cell therapy at Mesoblast. And I've been a consultant, also, since 2007.

So really what I was asked to speak about are particularly challenges faced by small companies, and maybe I should have now moved all of those other four things, because all of these bullet points really are the challenges that small companies face. And I think the most important thing coming from FDA or coming from academia is really to recognize that as we try to fund these pivotal or early clinical programs, often, we do this without any revenue generation stream. So literally you're just spending investor money.

And you could imagine that when you're the person responsible

for these clinical programs, you need to have strong board of directors, senior management support, but you also have analysts in the industry who watch you and look at you and make recommendations about how well you're doing, and that's both for public or private or just your VC investor support.

And then as you're doing this, a 5 or a 7-year or a 10-year clinical program, also recruiting and maintaining experienced clinical teams and pre-clinical teams becomes challenging because that uncertainty of getting your C round of financing to continue is always something that people worry about when they're working in a small company.

And so really what we all look for, but I don't think we always feel we get, is enough FDA input, and we need equality of FDA input, because who do we communicate that back to? Board of directors, senior management, and all those people who either directly support us or who are critical of how we're performing.

So when we talk about assurance of clinical direction, I really think that the FDA and the company need to have a thorough understanding of what the company's intentions are for its clinical development program. And that would be what I would encourage every small company to do very early because clearly FDA's expectations may not match exactly what the company's intentions are or financial capabilities are in terms of moving a program forward.

And I like to think about this as in the early 2000s in drug

development and biotech, the key thing everybody asked you was: Do you have an SPA? And the reality is, an SPA, a special protocol assessment, is only part of what we need as a small company from the FDA. So I used the concept of SPA^2 or SPA^3 because it's much more than just what your clinical trial design is. It's that whole regulatory submission package that you envision today that you'll need 3, 5, 7 years down the road that you think will support the approval of your device, biologic or drug, and keeping that dialogue with FDA going so that as new situations arise or as new outcome instruments are adopted, you can keep ahead of the curve, because remember, you have to go back to your board and back to your investors and convince them maybe that you need two more years to finish your program, because as we all know, this is a dynamic industry.

So I do emphasize that it's really that entire PMA package that we need feedback from FDA on. And, again, that's the assurance for everybody, but also developing an accurate timeline for funding. So we're all stressed by that of how much -- what's our burn rate. But we also need to project accurately into the future about what the funding timeline is going to look like.

So I've heard about we're not going to talk about MCIDs anymore, but for the time being, I'm still going to talk about them because I really want to underscore something that we haven't talked about yet. But from an industry perspective, there's a big difference between scientific,

clinical, and corporate goals with respect to patient responses. And, remember, when you're standing in the corporate realm, reimbursement needs -- and I say drive -- that might be too strong a word -- but certainly, reimbursement potential needs to be factored in. And it may, in fact, not be relevant to what the MID is. It may be something else. And I'm going to give you an example in a minute. And often, scientific and clinical decisions on endpoints are not correct, and I think we all know that, or not sensitive enough.

So we worked on an injectable peptide with a desired label for acceleration of fracture healing. At the academy, there was a fracture-healing symposium, when the panel said that a 25 reduction in time to fracture healing is clinically important.

So we took this in a reimbursement standpoint, and we spent a year running a statistical approach called conjoint analysis. This is how pharma and biotech develop pricing models for their therapeutic interventions. And this was done with 300 orthopedic surgeons; that's how we developed the questionnaire. And then we presented that to 250 medical directors here in the U.S.

And what we came up with was if there was a 25% reduction in time-to-healing, as the panel felt would be clinically meaningful, that would yield a $500 reimbursement. A 40% reduction in time-to-healing would lead a $2100 reimbursement. So, again, thinking about how was your device, how

was your drug, how was your biologic going to get paid for, and what is the clinical evidence you have to generate to ensure years down the road that you will actually have the ability to recoup that investment; this is so important today in small companies, medium companies, and I would say even large companies. It's just the scale of investment is different, say, in a Pfizer compared to OrthoLogics, who had 1.25 employees.

We also showed here no statistical sensitivity for gender, age, or BMI. But, remember, that's specific for acceleration of fracture healing.

So we all agree that patient-reported outcomes are critically important. And remember the way we like to look at it is a new device, drug, or biologic has to have a important clinical benefit. And we also know that imaging alone cannot be a surrogate for clinical benefit. And I can just tell you that based on experience in orthopedics, years ago, you could use imaging as your primary endpoint. And I think now, in the spine literature and the osteoarthritis literature, we know that imaging is a component of an endpoint. But it's really clinical benefit. And unfortunately or fortunately, depending on whether you're small and innovative or you're just small and a "me too" product, comparative outcomes are driving CMS decision making about reimbursement. I think we all appreciate that private payers generally follow CMS guidance. And also, orthopedic surgeons also are starting to look at clinical benefit.

So I think, as you've heard from other people, but maybe from a

small company perspective, we're all about composite endpoints. We think that's the trend today for approval, but certainly for the future. And if you think about therapeutic development or therapeutic device development, these are the new norm combining imaging function, for example, reduction in pain, and quality of life, obviously, also with safety if it's a first-of-its-kind technology.

And so just as an example, this is a recent FDA approval of the coflex device from Paradigm Spine. This is the label. And all I wanted just to show you is that here's an example of today, just a month ago, an FDA approval of a device that used a composite clinical success. So basically it had to be on the Oswestry Disability Index, a spine outcome instrument; an improvement of at least 15 points at 24 months compared to baseline; there could be no difference in revision surgeries or supplemental fixation; there couldn't be any device-related differences in treatment-emergent adverse events; and then the number of epidural steroid injections allowed, there could be no difference in coflex versus fusion.

So this is really the way we're all thinking about this. And I'm just showing you the one slide showing from week 6 through month 24, the Oswestry improving both in fusion patients and coflex patients, and you can look at this -- the statistics on the right. But I think, again, that was one of those composite measurements for approval.

The last slide is really talking about some work that I've helped

Mesoblast with. This is actually a study I designed. It's an early Phase IIA study of an allogeneic cell therapy product. But, again, our idea is a composite endpoint here. We think that if we're going to talk about cartilage repair or regeneration, we're going to have to have imaging, so we're doing total cartilage volume as well as T2 mapping to get an assessment of that cartilage quality. Obviously, since these are cells that are only in clinical evaluation right now, safety is an important part of that composite endpoint. This is an old set, so we're using KOOS, actually, the Marx Activity Index as well as SF-36 for function and quality of life as well as just a straight VAS pain.

I think the challenge here is: How do we set the thresholds for what all of these different measurements would be? And how do we weigh them in a composite endpoint? And this is really where I think a company has to be engaging FDA and having ongoing discussions with FDA about this, because you all appreciate that once you enter that pivotal trial, it's hard now to change this and the statistical analysis plan in an ongoing way. It's not easy to do adaptive clinical trials when you're talking about 3 months between follow-up visits, or 6 months between follow-up visits, 2-year follow-up or 3-year follow-up. It's not easy to design an adaptive clinical trial. It's easy to do that for blood pressure studies but not for the types of studies we're talking about here. But, again, maybe more creative trial designs will be helpful in the future.

So I think just to summarize -- I hope I kept my time -- engaging

FDA early, often, and having a good rapport with FDA is critical for small companies, and I would say every company today. MCIDs and that determination is a complex issue, as we've heard about all day today. But certainly reimbursement potential is a key component that we as a company always have to think about. And we really like composite endpoints because composite endpoints in early stage clinical trials can help you better focus on what are the sensitive endpoints and what are the endpoints that are going to be able to show that clinical benefit.

So thank you very much.

(Applause.)

DR. MIRZA: Great. Thank you, Dr. Ryaby. I was getting kind of worried. We were losing letters off of MCID. And then it just disappeared. And now it's back. And so let's see where the pendulum swings.

So here's Dr. Greg Brown from University of Minnesota, who actually helped -- I think he designed the Minnesota Total Joint Registry. And he's going to talk to us today and share his perspective.

Thank you.

DR. BROWN: Unfortunately, I can't claim to have done that. That was done with Terry Joy (ph.) and the Health E-System before I was -- well, I was still a resident, so -- but thanks for the thought.

So I think this is going to take a slightly different tangent from where we've been. I'm coming at this as an active clinician. I am not an

academician. I am an associate chief of surgery at a large clinical -- or large community hospital, essentially, Methodist Hospital in St. Louis Park, Minnesota. I left the university a few years ago. I still have an academic appointment there. I'm also chief medical officer for KareOutcomes, which is an outcomes management tool that we're using at Methodist Hospital.

Disclosures are KareMetrics or KareOutcomes, I'm chief medical officer. I have institutional support for a hip fracture outcomes study from Smith & Nephew, and I have two speaker honorariums, so noted.

The issue from my perspective and for the patient's perspective is that the cost of care being shifted to patients. And so last year, there were 27.7 million working-age people enrolled in high-deductible health plans. The fine is over $1,000 deductible per individual; $2,000 with families. If you look at where industry is going, 70% of large companies surveyed saying they're offering high-deductible insurance in 2013, and nearly a fifth of them said that's going to be their only option. So they're shifting things to the employees or our patients.

Several years ago, Medicare opened the door on having patients pay for their implants. So this was in cataracts surgery for lenses; they would allow patients to get a possibly higher-quality lens, but they had to pay the difference out-of-pocket. Before, that was never allowed. It's not currently allowed in implants for orthopedic surgery, but in an effort to control costs, I think that this is certainly coming down the road.

The point of these last three slides have been that patients have more and more skin in the game, and it's the patients that need to be involved in their own healthcare decisions, and we need to give them the information that they need to help make better decisions.

There's a program that was from the American Board of Internal Medicine, Choosing Wisely. It's to help reduce waste in U.S. healthcare. I think in orthopedics, it's not necessarily waste in the sense that the patients don't need the operations. Sometimes I think it's often premature. Sometimes I think the expectations are out of line with what we can deliver. And we can't measure that or understand that, and we don't have the data to tell them what their expected outcome should be.

Park Nicollet Health Services is a large integrated healthcare system. This is where I work. Revenues last year were $1.2 billion. We were one of the original Medicare pay-for-performance project sites with the Physician Group Practice Demonstration project. We're a pioneer accountable care organization. We recently announced a merger with one of the other large groups in town, HealthPartners. They are both a large health insurer and a medical group.

What I have done as associate chief of surgery in Outcomes in this hospital for the last three years is go through every single surgical subspecialty and select quality of life, which we've selected EQ-5D, and we can talk about it in the discussion area, and also then disease or procedure-

specific patient-reported outcome measures. I've vetted those with each of the departments, sat down, had discussions, and we've picked them. And primarily, our criteria was pragmatic and say -- we thought the biggest issue was respondent burden. So if we had a validated measure that had fewer questions, essentially, that's where we went. But we did it in all these different areas.

The reason we're doing this is I think what's driving a lot of the issues is trust. There's a lack of trust. There's a lack of trust between patients and doctors. There's a lack of trust between doctors and insurance companies. There's a lack of trust between insurance companies and patients.

And how can we counter that? Well, I think part of the way you counter that is through transparency. Here's an example. I spent the first 7 years of my practice going to a small town in west central Minnesota. When I came to Park Nicollet, one of my former primary care docs out in Wilmer sent in this patient. He's a retired veterinarian. His hobbies were hunting and fishing. He was sent in for a second opinion. Another orthopedic surgeon suggested he needed a total knee replacement. He's certainly having left medial knee pain, and you know, he meets radiographic criteria for a knee replacement. He's certainly got Grade IV changes on Kellgren and Lawrence scale.

The issue was, you know, I said, well, are you hunting and

fishing right now? He said yes. I said, well, then why do you need a knee replacement? I don't. Bye. You know, five-minute consult. That's okay. Well, the issue is I think what's happened to the definition of evidence-based medicine. And this is from David Sackett 15 years ago. And, you know, right now when you think evidence-based medicine, what randomized controlled trial do you have? Well, the problem with randomized controlled trials is they're in populations and not individuals, and we have individual patients when we're clinicians sitting in front of us, not a population. And so it's the current best evidence in making decisions about the care of individual patients. I think there's the evidence.

But the issue that often, to me, gets lost is the last part. And that's the thoughtful identification and compassionate use of individual patients' predicaments, rights, and preferences. You can never look at patient preferences in a randomized controlled trial. And so you have to be able to synthesize, you know, both the evidence and the patients' needs.

And there's the example. You know, he certainly met radiographic criteria for a total knee, but quality of life-wise, expectation-wise, he wasn't ready for a total knee yet.

Sorry. And so I think here's where evidence-based medicine is going. It's a prediction. I think we went from pre-evidence-based medicine to what I will call EBM 1.0 where we're looking in the therapeutic domain, and randomized controlled trials and meta-analyses. I think we need to shift into

the prognostic domain, where a observational outcome registry that has 80% follow-up is Level I evidence. But that's where we can actually now counsel patients with prognostic factors of what to expect with a surgery such as a knee or hip replacement.

Ultimately, I think we need to get to the fourth column, which is the economic one, where we start looking at value. It's really value-based healthcare, because that's what the patients are paying for when they've got all the skin in the game.

So this was a book that came out several years ago. I don't know if you read it or not. It's a little bit of a tome. I think the key is just the subtitle: "Creating Value-Based Competition on Results." Well, here's my response to that, I shouldn't, but I'm going to have, the garbage. Well, the data out there, if you're a patient on whether you should have an operation or not, is garbage. I mean, a quality measure? So if I stop my antibiotics after a knee replacement at 23 hours and 59 minutes, I meet the criteria; if I stop it at 24 hours and 1 minute, I fail. Does anybody think that affects the patient, the outcome, the care that I deliver to my patient? No.

So, you know, I can meet every criteria because I'm in a hospital with a great infrastructure and has everything set up with electronic order sets and an EMR. And I meet all the quality criteria. Every one of my patients can limp for the rest of their life and be maimed, but on the internet on, you know, CMS Compare, our hospital has a great quality rating; everybody thinks

I'm a great doc. Well, you know, the point is the evidence out there isn't what it needs to be.

So to look at value, we need to look at not just total cost of care. We need to look at the quality-adjusted life-years. And, you know, here's briefly what a quality-adjusted life-year is. It's come up a few times this morning. Essentially, if you have one treatment program and another treatment program, it's the difference between the two curves. So say, for example, you do a knee replacement. You have a 15% improvement in quality of life. That knee replacement lasts 20 years. That would be .15 times 20 years, or 3 quality-adjusted life-years. Then you divide that by the 3 by the total cost of care to get how many dollars per quality-adjusted life-year.

So I think you can use that information now to try and help our patients. So, you know, there's a simple decision making tool here, you know? Look at risk-adjusted outcomes for a specific patients; they could be poor, good, or excellent. Look at risk of adverse events. It can be a single adverse event like an infection or it can be a composite of adverse events; infection, venous thromboembolic event, 30-day reoperation, 30-readmission.

Well, right now, we don't have the data to do that. So how can we get there? Well, this was an article that just came out this year looking at what do we need for outcomes reporting. Well, they tell us the clinicians' recommended metrics or the mean change in an outcomes score, a patient-reported outcomes score, and the proportion achieving a minimal clinically

important difference -- sorry, I made my slides before we -- I got my lesson this morning. For patients, we recommend the proportion achieving a minimum clinically important difference.

And so that's what patients need to know. That's what they care about when they're sitting in front of you in the office and want to know, "Should I have this operation or not?"

Also, you know, same idea for meta-analyses. We should get away from other measures, and we should be looking at minimum clinically important difference or the smallest difference patients experience is important as a way to report meta-analyses and look at -- do them.

And so I made this slide after I have two of the authors here, the two that were the et al. Sorry. I apologize for the inclusion. But Dr. Sloan spoke this morning and Dr. Wyrwich just spoke. And, you know, I'm an engineer also by training, so I'm very pragmatic and expedient. And this works very well as a way from a clinical perspective if I've got a set of outcome data and I need to analyze that for patients. And we'll give an example.

So we can now grade outcomes for an individual operation with a patient-reported preop and postop outcome data. If the change was less than 1 MCID, it's a poor outcome. If it was between 1 to 2 MCIDs, it's a good outcome. If it's greater than 2, it was an excellent outcome.

So how do we operationalize that? Well, here's some lumbar

discectomy data using Oswestry patient-reported outcome. I picked four patients and then gave the summary data, but you can see that on the change, the mean change was -35 points. And the standard deviation was 10 or 20, I'm sorry. So the MCID is going to be half of that, or 10.2. So now I go back and divide the change by the MCID of 10.2, and I report the outcomes now as MCIDs. So -4, that's significant; -1.7, -2.5; I have one where it's only -.98, so at least according to this approach, that's a failure.

What I then can do is I can go through all the patients that I have in this, you know, outcome registry on lumbar discectomies, and I can look at what are the MCIDs for Oswestry, back pain, or leg pain. I can look at the successes. I can look at the failures. And I can now tell you success rates.

And where that's important is when you start looking at subgroups. So let's look at Oswestry failure rates based on males or females and age, less than 50, greater than 50. And what you find is that there's a almost five-times higher relative risk for males over 50 compared to males under 50. So the surgeon whose data this is that I analyzed for them now counsels males over 50 very differently than he did before looking at his own data. I mean, this is where we're supposed to do where we talk about looking at your own practice, looking at your own data, and improving the quality of the care that we're giving. And the way we're doing it is through looking at the minimum clinically important differences for individual patients.

The other thing we need to do is we need to risk-adjust the

data. I would suggest that administrative data, such as diabetes, tobacco use, isn't nearly as good as hemoglobin A1C, pack years, BMI, GFR, which are all continuous variables. I'm not a biostatistician, but I think all the biostatisticians would agree.

Now, we can go back here and see how we can populate this decision -- shared decision making tool that we use with patients. So we can look at for an individual patient, what's your predicted outcome based on demographics, comorbidities, and diagnosis; what's your risk of an adverse event, based on those; and where do you fall? If you're a 5, you're an excellent candidate and have a low likelihood of a adverse event. You know, you would probably want to proceed with surgery. If you're a 1, you know, you expect a poor outcome, and a high risk of adverse event, then I think patients don't -- most patients don't want to have an operation just to have an operation. They want it because they're going to be better. And if you don't think they're going to be better with some certainty, they're going to listen.

So now we can go back and now we can -- I don't think the competition -- I don't think I do a knee replacement better or worse than the average orthopedist, but I think the key is how we select patients and manage things, particularly their expectation. And so one example is satisfaction should be the outcome minus their expectation. Well, how do you do that? Well, you know, if I say what do you want for a quality of life improvement,

that's really hard. You know, if I say what do you want for a knee function improvement, they look at me like what are you talking about. Well, if you give them concrete things like level of activity, you know, and they have a scale that they're looking at, or a Visual Analogue pain scale, I want to go from an 8 to a 3, you know, that's very concrete. And now you can measure their expectation preoperatively with these measurements and measure them at follow-up appointments, you know, 3 months or a year that we're doing for total hips and knees, and you can actually measure whether you did meet their expectation or not.

One thing just, you know, I find every time I talk to orthopedic surgeons, I find this analogy very good. If anybody hasn't read it, it's an excellent book, *The Social Animal*. But "clocks are neat, orderly systems that can be defined and evaluated using reductive methodologies. You can take apart a clock, measure the pieces, and see how they fit together. Clouds are irregular, dynamic, and idiosyncratic. It's hard to study a cloud because they change from second to second. They can be best described by narrative, not numbers."

Well, our patients need better ways to make decisions, and all these techniques that we're talking about and the statistical processes are very precise, and you can have, you know, four significant digits. But the point is we need something at least with one significant digit to help our patients make better decisions. And this is a way to go at it.

The other thing, clinicians are overwhelmed by all the outcome tools, you know? There's books on them. And how do you pick them? Well, I think for joints, you know, I think some of the international registries have already done this for us, you know? They looked at EQ-5D, and they looked at the Oxford ones. We also have a measure in Minnesota, Minnesota Community Measurement that we have just adopted -- I was on that working group -- looking at total knee replacements. The other thing is researchers looked at it for femoral neck fractures, it's EQ-5D; inotropes, EQ-5D, and subtropes, EQ-5D. So I think the pragmatic approach is to do EQ-5D, Oxford or Lower Extremity Activity Scale, and VAS, which is only 20 questions versus 60 question if you do SF-36 and WOMAC.

And here's just a brief slide showing how many dollars does it cost for a quality-adjusted life-year. Orthopedists should not be worried about competing in policy discussions about how much improvement we give to our patients. We do very well with what we do for patients.

And so go back to our case: Building trust through transparency. Who do you think this guy came back to see? Well, he came back to see me. He has his uni, he's very happy, he's still hunting and fishing.

So conclusions, we need to standardize patient-reported outcomes so that we can collect this data and help our patients, you know, make better decisions, choose wisely, and build trust with them with the data that they have.

Thanks.

(Applause.)

DR. MIRZA: Thanks, Dr. Brown.

Now we'll have Janice Hogan from Hogan Lovells, who will give her perspective as "Role of the Consultant."

Thank you.

She is a lawyer, but we still welcome her well.

(Laughter.)

MS. HOGAN: I can understand if you're all wondering why would anyone go to an outside law firm or really any consultant for help on this topic. But I think maybe what I can bring to the equation today is the topics that people come to us for help on over and over again, probably point to some areas where a meeting like this should be really helpful in giving people better clarity. And I tried to leverage my example in several dozen orthopedic studies working towards FDA approval over the past 20 years and pick out some issues related to PROs and MCIDs, although I've learned today that I should stop using that terminology, you know, to illustrate some things that companies face over and over again.

And I want to commend the meeting organizers and the FDA for hosting us, because I would say, in my experience, this is a really big challenge for companies. In some areas, where it's very well established what instruments are tried and true and what level of improvement defines a

responder, those are the easy cases. But there a lot of more difficult cases.

So what are some topics that people routinely ask me about? And I should say from the get-go that unlike many of the people here, my experience is really limited to company-sponsored, industry-sponsored studies, and so cost and time are always important factors that influence how those studies can be run.

But some of the topics we've been talking about all day are the same themes we talk about with companies repeatedly: What are the best instruments to select? Which ones are regarded as adequately validated? And as was discussed earlier, a questionnaire or an instrument may be validated, but is it validated enough and in the right population and at the right time points? What is the threshold for success that would define a responder? How can they interact best with FDA to work out these issues? And how do they balance the needs of FDA requirements and not only reimbursement requirements but, as we've talked about already, all of the various constituencies that ultimately will have to accept and use the data?

So when companies come to us, very often, they will essentially commission us to survey the entire body of literature and review dozens or hundreds of papers and distill for them all of the instruments that have been used, which ones have worked the best, what is our experience. If we're very fortunate, there's an FDA guidance document that already details what instruments are acceptable to the FDA. But oftentimes there may not be, not

in every therapeutic area in orthopedics. And so I think it's a great effort to define what PROs are recognized and validated, because oftentimes companies have to spend a substantial amount of time and money just to figure out what is the universe.

Then we look for, obviously, what instruments are adequately sensitive to change considering the target population that we're looking at. What is accepted by FDA? As I said, if there's a guidance document and it's well-known or there are six other products of the same type that have already been approved, then we don't have a very hard job here. But oftentimes that is not the case.

And then once we get past the basic threshold of what instruments are acceptable and validated, then we look at other factors such as what do clinicians feel comfortable with, which instruments are going to be the easiest for patients to complete, the most repeatable, the most reliable, and what will payers accept.

And fortunately or unfortunately, oftentimes where I come into the puzzle or we come into the puzzle is companies are looking at these factors and they're getting a lot of disagreement among the various sources of information about what would be the best instrument.

So some of the key problems I'm not going to spend time on because we've already discussed them today.

Is an instrument considered validated by the FDA by other third

parties? It may be validated, but was it validated sufficiently and correctly in the right population at the right time point for the right disease severity?

Is it validated in all of the different languages? This is becoming increasingly common, where companies, because clinical trials are very expensive, they need to really leverage them to get the most out of every study. They may want to use a study multinationally to get approval in multiple markets, and so what to do if the PRO that they want to use is not validated in all the relevant languages.

Although it is theoretically possible for a sponsor to conduct its own validation, this doesn't work in all cases and doesn't address all the concerns. If a company has to go out and validate its own instrument, this would raise a lot of concern, generally, because that means it's new enough that we don't know if clinicians would accept it, we don't know if payers would accept it. They could do an awful lot of work and spend a lot of time and money on the validation not to know in the end whether this will really yield something productive for them.

As we've been talking about all day, the threshold for what constitutes a meaningful degree of change such that we would call someone a success or a responder can be very difficult to establish especially for things that are innovative. So suppose we have -- current medical gives us only one very invasive method of treating a disease and a company comes along with something that's much less invasive; we don't know often how or whether we

should adjust the threshold for success that would define a responder to take into account a lower or higher level of risk. And so these are areas where even if there's quite a large body of literature, we usually don't have all the answers.

Another area where we've had a tough time figuring out what to do is in situations where we have an adjunctive treatment or a combination of treatments. Again, this is increasingly common. So suppose, hypothetically, you have a surgical treatment and you know that that achieves a pretty effective improvement on a particular PRO. And you have a product that would enhance the outcome of that primary treatment. What threshold for response do you define for the adjunct? Usually it can't be the whole MCID or MID because the product is only an adjunct. So this is very difficult because most of the time, you will not find papers in the literature that establish what an MID would be for an adjunctive treatment as opposed to a primary treatment.

As was pointed out previously, one of our biggest challenges is trying to get all of the parties to agree, which really doesn't even happen because it can't happen because all of the constituencies are never in the same room together. So, typically, companies will try to work things out with regulatory authorities first, with input from investigators and scientific advisors. Only recently have we really had the opportunity to do that kind of front-end work with payers. And to try and make it all come together is

extremely challenging.

Sometimes in the past what has been done, and frankly, it's been a mistake in some cases, is to try and layer on multiple instruments to please everyone. So if we were trying to do a study to satisfy European authorities, U.S. regulatory authorities, private payers, and clinicians, we end up using four or five or six different measurement tools, and this doesn't really work out well. And so a red flag for me would be if we're giving the patient a booklet of 20 pages of questionnaires to fill out six times over a longitudinal study, this probably isn't going to work out happily for everybody, or at least we're going to have a lot of potential inconsistency and incompleteness in the dataset that we'll have to deal with.

Just to give an example, and there are many of these, I think out of all the questionnaires we've talked about today, I can remember having at least a vigorous debate if not a true downright argument about which would be the best in pretty much every joint in the body that we covered, with the possible exception of the hip, where the Harris hip score has been the go-to measure for a long time.

But here's an example from a few years ago of a tool called the Zurich Claudication Questionnaire, which at the time it was first used in an FDA-regulated study, to my knowledge, it was regarded as a validated outcomes instrument for spinal stenosis. An MCID had been defined in the literature. Although better defined perhaps today, there was validation. And

it depends, you know? As we said, how much validation is enough? It has been accepted by FDA in multiple studies, but at least originally, it was not very familiar at all to clinicians or to payers. And a question that arose in the FDA Advisory Panel Meeting was that the questionnaire was validated only at 6 months, not for the longer-term endpoints that were ultimately required by the FDA.

And even though there was literature establishing the MCID, which coincidentally was .5, in the Advisory Panel deliberations, the question at the bottom of this slide was raised. And I think this is a pervasive question that we've been talking about over and over again today. And I think in this case, although this ultimately was worked out, the lack of familiarity of the questionnaire even though it was validated and an MCID had been established, still that wasn't enough to completely smooth the path.

Just a few other considerations. One question we are asked all the time: Should the same PRO be required for all products in the same class or for the same target population even if there are multiple validated instruments? And certainly from the regulatory perspective, if everyone in a particular product class uses the same PRO, it's convenient. And it's convenient not only for regulators but probably for doctors and patients as well to be able to cross-compare similar products. However, for an individual sponsor, this may not be what best suits their needs.

And I'll take the example from earlier today. If somebody

develops a new and creative instrument that looks at hand function that would be particularly suited to younger patients, a sponsor whose product is designed for younger patients might prefer to use that even if that has not been common in the past. But there is a lot of inertia here for multiple reasons, again, because of those multiple audiences that we are targeting.

Many of the issues we've talked about today can be worked out with FDA through pre-IDE interaction. And this can be very successful especially on instrument selection, whether an instrument is regarded as validated, what should the threshold for success be. But one of the things that the pre-IDE process can't really resolve for us is this issue of disagreement between multiple audiences of which instrument or instruments are best. And all the parties also may not agree on what constitutes success; a payer's interest, particularly in what constitutes success from their perspective, may not match up with a regulatory standard for safety and effectiveness.

So, in conclusion, I think it would be really helpful if out of today's proceedings, some of the key challenges for sponsors could be discussed further: Validation of new instruments or of existing instruments in additional populations is a significant burden for companies, something that most companies would not take on to create and validate a new instrument because of the time and cost. Establishing thresholds for success particularly in some of those tough situations I mentioned, adjunctive treatments, for

example, and this issue of meeting the needs of all audiences.

Thank you.

(Applause.)

DR. MIRZA: Thank you. And I just wanted to note that Janice was three minutes early. So maybe we should have more lawyers.

Now I'll ask David Appleby from Smith & Nephew to provide an "Industry Perspective on Patient Factors that Affect Outcomes."

MR. APPLEBY: Well, good afternoon. One of the nice things about being one of the last speakers is pretty much everything's been answered already, so --

(Laughter.)

MR. APPLEBY: Not really.

So I was asked to give our perspective on patient factors that affect outcomes. And so these are just some of the things that I'd like to touch upon in the next 15 minutes. And even though we've -- a lot of people have talked about the minimum clinically important difference today, I do want to throw in my two cents, in particular, in the orthopedic surgery realm.

What I would say is, you know, there's not general agreement, okay? And so even, like, early this morning, talking about, you know, a statistical method for, you know, defining certain things, the anchoring and all of that, when we actually go and talk to orthopedic surgeons, you know, does this really have meaning to them, all right? What I would say is, in my

experience, if we talk about a hip replacement, you know, most surgeons understand the Harris hip score, and if you just get a status at a certain time and a certain number, that has meaning to them, you know? If you have a 90 or greater, that's an excellent outcome.

The other thing, just looking at some of the scores that we use on the arthroscopy side -- in particular for us, we use the KOOS a lot -- if you go back to one of the original articles on the KOOS, you know, those authors suggested that an 8-point difference was the minimum important difference. But there's a million outcome measures. You know, I didn't even want to go into listing these.

And, in particular, when we start -- I think a lot of the outcome measures that we've talked about today really were more on the adult joint reconstruction side. When we start going into the arthroscopy side, I think it almost gets even a little bit messier. There's a lot more consistency in the outcomes being reported in total joint replacement surgeries. If you look at -- you know, if you do a literature search on ACL reconstruction, for example, you know, nobody is reporting the same things. So it just gets really crazy and difficult for us, really.

So when we're talking about, you know, what's an important difference -- and I think -- and I apologize -- I forget your name -- the lawyer kind of hit upon this probably much more eloquently than me. You know, we go and talk to all of our different stakeholders, you know? We do clinical

studies for a lot of different reasons, all right? And we have lots of different people that we need to get input from. So the regulators, you know. We also want to get the word out, so we need to make sure that what we have in our studies is something that we have the ability to get published. We want surgeons but, you know, two different types of surgeons. We want surgeons that are clinicians, and we want surgeons that are researchers. And obviously, we want statistician input. So when we're talking about what's a meaningful improvement that we want to see in a particular outcome, we have to go and talk to everyone.

So anyway, you know, moving on from that into some of the stuff that we do to account for patient factors, you know, the big concern that we have is confounding. And I'm using confounding in a more broad sense, to include such things as effect measure, modification, and other things, and that's for all the statisticians and/or epidemiologists out there.

But really, you know, we're concerned about the patient factors because you have this mixing of effects. When we're doing a study and we're comparing treatments, you know, what we want to know is, you know, what are the outcomes associated with that treatment, not these other things that may mask it, all right? So when we talk about mixing of effects, those are some of the things where -- I'll go to, like, one of our cartilage studies, actually. You know, smokers have bad cartilage. And in particular, when we're talking about articular cartilage defects of the knee, you know, that's

something -- smokers just have worse results, they start off worse, and everything else. And that also -- you know, we have baseline inequality.

So one of the things that we get concerned about is if your two treatment groups aren't balanced, right? And so, you know, you try to randomize and other things, but sometimes you just end up with two groups that don't look the same. And so those can mask the actual treatment effects that you have.

All right. So the way that we handle, you know, controlling confounding is you do it both in your design and in your analysis. All right. So in the design side, you know, one of the things that we all do is we develop this inclusion/exclusion criteria. And, again, we do these things for lots of different reasons. In our arthroscopy studies, for example, for a rotator cuff, you know, we may have an upper age limit because we know older patients with a rotator cuff tear, they just don't heal in the same way that a younger person does. So we'll try to control, you know, some of those factors in the patients that we allow in the study. You know, the other thing is your control groups, your randomization -- a lot has been talked about a stratified randomization, which I think is a great way to do it.

Now, on the analysis side, I have listed stratification, but this is a little bit different. And if you look at the orthopedic literature, both in adult joint reconstruction and arthroscopy, it's everywhere, all right? It's probably not my preferred method, but you're going to see it. So you can take your

results, stratify by sex, for example, and then kind of look at your results, you know, between males and females and see if there's a difference. And that's okay. I mean, there's nothing inherently wrong with that, I don't believe, but it doesn't give us, you know, the best picture.

So what I like is a multivariate analysis, because you want to be able to control for a bunch of these things. But if you go through the orthopedic literature, it's mostly univariate analyses that are reported and presented.

So I'm just going to give you two very brief examples from stuff that we have done just to kind of highlight a couple of these issues. So for our particular product that we were looking at for articular cartilage defects of the knee, you know, some of the confounding is concomitant surgical procedures. So patients that have, you know, small, isolated cartilage defects of the knee very often have ligament involvement and meniscus involvement. And for a lot of reasons, going back to that original slide when we go and talk to everyone, we wanted those things to be excluded and just have an isolated cartilage defect, you know?

So this on the design side; your inclusion/exclusion criteria trying to control some of the confounding. You know, we had a microfracture control group, all right, so that's put out as the gold standard, and you know, how close are we to the gold standard. And then, you know, randomization is really your best bet and a sufficiently large population in trying to get rid of

inequality between your groups. So that's just, you know, an example of controlling some of the confounding in your design.

And then on your analysis -- and this one I'm going to punt a little bit, you know, because with the adult joint reconstruction, almost everything that's out there is really on implant survivorship, you know, so everything is either being reported revision rates or survivorship, you know? So the good thing is, you know, if you're looking at survivorship, you can do a multivariate analysis. You can look at, you know, hazard ratios for sex, height, weight, you know, any of those types of things, and you can get, you know, what the excess hazard is from those individual effects, you know? So that's what I like about that.

And in this one, there was a hip score and patient satisfaction score. And in this particular product, you know, we did go and stratify by the diagnostic indication for the procedure. So that was one of those simple, here's the diagnosis, you know, the three different diagnoses, and this is what the scores look like, so --

And that is it.

(Applause.)

DR. TURKELSON: So the floor is now open for questions should anybody have any.

Dr. Jevsevar?

DR. JEVSEVAR: Dave Jevsevar. I'm from Utah, Intermountain

Healthcare.

Greg, this one's for you for the patient expectations and how you can validate one of these measures or these measures for patient expectations. Let me give you a personal example. So I had a drop foot, and I had spine surgery in the morning. And I go in for my preop. And I'm laying there in the gurney. And everybody is all nervous because I'm a doctor. And the nurse asks me. She goes, "So what are your expectations for pain after surgery?" And I said, "Well, I want zero pain." And she goes, "That's interesting. You're the first person that's ever said zero."

So we looked at the next month's worth, which was 208 patients, and she was right. Not one patient had zero as their expectation for pain. Functionally, I would have had a different result. And even now, it depends on what I'm doing as to how my results are. If I'm skiing, I feel great and everything's perfect. If I travel across the country to Washington D.C. in a plane and sit all day, things aren't as perfect.

So how do you validate these expectations from these types of patient-reported outcomes?

DR. BROWN: You're describing the classic problem with any patient-reported outcome. I mean, it's do you do it that day, do you do it over the past 2 weeks, do you do it at any time, and that's just the nature of the clocks and clouds comments of, you know, there isn't a precise measurement, and the issue is we're measuring as best we can with an

instrument that has been validated, you know, in other ways. But it's, you know, can you state for an individual patient is it validated at that measurement, period? No. That's why it's a statistical process, I mean, that you're -- or probability that looking at a group of patients over time, that they are sensitive to change, that they're, you know, internally valid and those sorts of things.

DR. JEVSEVAR: So when you talk about being able to better inform -- shared decision making, how do you take the MID and turn it into something you can tell a patient, to say this is what, you know, this is what the difference is that we look for, but what does that mean to you. And say you have one patient who is very active and plays pick-up ball every day but needs a knee replacement, yet you have another patient that's in a wheelchair and needs a knee replacement. How do you turn that statistical number or those numbers into something you can actually give to patients that makes it meaningful to them?

DR. BROWN: I think I'm going to answer a different question, and to me, ideal healthcare requires three components. It requires evidence; I think we have good evidence for, you know, the right patient that a lumbar discectomy is very helpful. I think it requires technical skill on the surgeon's part to be able to perform the operation well. And that is a psychomotor skill, so we're not all created equal. And I think the third part is really an emotional intelligence issue, you know, how empathetic are you, how much time do you

spend with that patient trying to understand what they are. And so you know, getting a measurement here of what do they expect is just that. It's a one-time deal, and you know, is that -- like I say, you can't validate a single measurement in time as saying that's valid. But you can say, you know, with 100 measurements, you know, 83 of them are valid, you know?

And so I think that's the skill in how we deliver healthcare is being able to have that emotional intelligence to hear patients and what they're expecting. I mean, one example of -- you know, it's the classic case where you get caught doing an operation that someone else had teed up.

It was a, you know, young male that had a bad OCD and had bone-on-bone in his medial compartment from the OCD and wanted a uni, unicompartmental knee arthroplasty. And another surgeon was all set to do it, and then the patient said, oh, I want this implant. And he's like, well, I don't know how to do that implant, you know? Greg, will you do it? Well, my first offer was how about I do it with you, you know, and that way you can continue to manage the patient. And he's like -- ultimately ended up, well, why don't you just do it.

And so we did it at a different facility than initially expected, with a different implant, and I never met the patient preoperatively. And, you know, in follow-up, the patient's like, well, I did this so I could go back and play basketball. I was like, well, if I had seen this patient ahead of time, I would have never said, you know, your expectation of going back and playing

basketball after a unicompartmental knee arthroplasty is appropriate.

And so, you know, that's the part of medicine that's very soft. And it's hard to measure. And what you can do is something like this. And, you know, right now, we have small numbers where we're collecting this on total hips and knees. You know, for my group, the average minimum clinically important difference is -- improvement on total knees is 2 MCIDs, and the standard deviation is 2 MCIDs. My personal data, the average improvement is 3 MCIDs with a standard deviation of 1 MCID. I clearly pick patients differently. Again, I don't think I'm a better surgeon than any of my partners. I think I pick patients differently. And that's the part that's really hard to tease out of this in helping patients make decisions.

DR. TURKELSON: Yes, ma'am?

DR. DAVIS: Dr. Ryaby, just point of clarification. When you talk about composite scores, initially, when you were talking, I got the impression you were talking about we needed to measure these different components and you were looking for criteria to be met on all the components in order to declare responder. But then at the tail end, I thought you were talking about an algorithm in waiting and you were looking for a single summary score. And I just was wondering, from industry's perspective, was it both or which one? I just wasn't clear.

UNIDENTIFIED SPEAKER: Sorry. My opinion would be for hips and knees that you pick one score like a Oxford, yeah, hip or knee score

because that's what you're doing the operation for. And health-related quality of life has so many other measures that that has to be a secondary outcome, in my opinion. And so that is a composite score because most of those have, you know, stiffness and pain and function as part of the instrument.

But what I meant by composite one is I was strictly referring to adverse events. So you could have a single adverse event, like I say, infection in a joint replacement, or you could have a composite score of adverse events. So you could look at infection, venous thromboembolic event, 30-day readmission, 30-day reoperation, whatever you think are the appropriate adverse events that are significant for that operation or procedure.

So -- oh.

DR. RYABY: Well, as a consultant, you're absolutely right. So that's what you learn as a consultant is that the people who hire you, you try to always tell them that they're right -- no, I think we don't have an answer yet. That's the point. I think when I think about OA or cartilage defect repair and I would want in my label that the repair cartilage or the regenerated cartilage is structurally sound or it's hyaline, I need to use an imaging tool that will give me that assessment non-invasively. So I think the issue is we all know you would want the pain to go away, we all know that you would want the function to improve, but we also want the best label. And now with these different measurements as a composite, how do you weigh the importance of

each of those, and how do you set the threshold for what that difference is that would be both meaningful clinically, as well as we do have to think about what would be reimbursable. And those don't always match. I tried to show you that with the one fracture-healing example. And so I don't think we have answers to any of this yet. But it would be nice to talk about some of these things tomorrow.

AUDIENCE MEMBER: Question for David Appleby. You brought up a point in your example of your cartilage repair study that I've grappled with with regards to study design. You decided to limit the lesions to pure cartilage lesions without any of the concomitant factors that are more common, actually, because a lot of those cartilage lesions are secondary to the ACL tear or the meniscal tear.

And so you've restricted yourself in two regards. One is, you're not really testing the real world because when you want -- when your product is approved, presumably it's good. You'd like it to be applied for all of those lesions if it would work. And, secondly, you've limited what you can say in your package insert because you've only tested it for that unique situation without concomitant problems in the knee.

So how do you -- I know where you came down on that, but why?

MR. APPLEBY: Yeah, that's an excellent question. That one, I did kind of cheat as well. That was not being done for the U.S. It was for a

product that we have only in Europe. And the reason why we did it in that particular way is from feedback from our European surgeons and some of the payers in Europe, okay, so that's the reason why we chose to do that. The other unintended, you know, unintended consequence of that, it took us two and a half years to enroll 150 patients. And to be quite honest, we were probably one of the -- you know, for these types of trials, probably one of the better enrollments that was happening in Europe.

You know, so it didn't affect our labeling for Europe because the way the CE mark is applied is different. And this was really being done for reimbursement and health economic reasons. So it didn't have the same level of importance in our labeling as you would here.

DR. TURKELSON: Dr. Keith?

DR. KEITH: I heard today that there's been a lot of change, and I just want to try to get some clarification. I heard the MCID is gone. I heard the MID replaced it. And I heard from two speakers that the MCID or MID is not relevant anymore. Could I have some clarification on that as we go forward into tomorrow's discussions?

DR. WYRWICH: Right. So first, MCID, minimum clinically important difference, Gord presented that in the early '80s, his group, the evidence-based medicine group at McMaster, had first coined that term. He first operationalized it in a 1989 study with Roman Jaeschke and Jules Singer. And they went to patients, the patient anchor, a very great deal better, a very

great deal worse. And they've repeated this process several times at McMaster University.

And in 1994, when they went to publish these important change thresholds, they realized that there wasn't anything clinically associated with what they were doing other than patients were in a clinic when they gave them the PROs and the anchor questions to answer. So the "clinically" part of it -- it's patient-reported outcomes, and it's patient empirically verified important change thresholds. And so they dropped the C. And from then on, we called it the MID, the minimally important difference.

The MID became interpreted -- and we talked about this in this room this morning, about the fact that are we talking about the difference between placebo and treatment group, or are we talking about the difference at the individual level. It became very difficult especially for FDA reviewers when they saw this MID to know exactly what the sponsor was talking about.

In 2006, February of 2006, we received the draft guidance on PRO use for label claims. And that had a section in it on responder definitions. That's at the individual level; did you meet this threshold or not. And it had a section on MID, which looked at the group mean difference in looking at trial results. And the wording there, although I'm not sure much of any of us really understood what it was saying, the wording there said that if you are using the term MID to talk about that group difference, that not only did the mean difference have to exceed that threshold you talked about, but

the confidence interval around that mean difference also had to exceed it in order for it to be eligible to be considered for a label claim.

In 2009, when we received the finalized PRO guidance, that MID term is totally removed. There is nothing in it about minimally important difference. It only talks about the responder threshold where we go back and look at individuals and report the percentage of individuals who have achieved an important change threshold. And that change threshold could be "getting better" or it could be "not getting as worse as you would expect to," perhaps in an oncology trial; and also, the use of the cumulative distribution function to not only talk about through responder definitions the percentage of individuals that achieved a certain change threshold but to also look across the entire spectrum of change and to be able to show that differentiation between treatment and placebo groups.

So MID and MCID are gone. What's replaced those have been responder thresholds, those who've met an important change threshold, which has been used throughout the day as that change threshold being the MCID, as well as the cumulative distribution function to be able to get the total picture. But the studies should be powered on that continuous variable, group mean change. And once p less than .05 or whatever threshold you're at based on what type of multiplicity you're dealing with, once that threshold's been achieved, then you want to go back and look at your percentage responders, and that can be reported in the label to help the clinicians to

understand what can be expected from the PRO as well as the cumulative distribution function.

DR. KEITH: I'm hoping for the sake of tomorrow's discussions that the statisticians in the group will huddle over a warm glass of wine tonight and tell us tomorrow whether or not we should take to our groups this new definition or to continue to look at what's being published tomorrow, you know, in the literature regarding MIDs, and how these are going to be reconciled for, you know, the guidance documents of the future.

MS. HOGAN: I think one thing that might be confusing to some people in the audience is probably just not careful terminology. So if you look at the last dozen or so orthopedic studies that have been approved by the FDA that use a PRO, there's a threshold usually that says to be a responder, a patient must achieve a 15-point change on whatever is the PRO of interest. And I know I'm guilty of calling that the MCID or the MID now, and I think what you're saying is that we should talk about that as a threshold for calling someone a responder?

DR. WYRWICH: Right, the responder definition or responder threshold.

MS. HOGAN: But what also I think has not come through, at least not in device studies that I've worked on, is the concept of the cumulative frequency distribution because as someone mentioned, oftentimes, in device studies for the FDA, we define a patient as a responder

only if they achieve some meaningful change on the PRO and they have no reoperation and sometimes they achieve an imaging element. I'm not clear myself, and maybe this is fodder for tomorrow, on how you would fit that kind of composite endpoint together with a cumulative frequency distribution approach. Another one for the statisticians.

DR. TURKELSON: Yes, ma'am?

DR. SUTER: Lisa Suter from Yale Outcomes Research Center and Evaluation.

And I think this sort of applies to all groups, and some people have mentioned the PROMIS measures or tools. But I'm just curious as to whether people have a perspective on either computer-adaptive technology in this forum or specifically the PROMIS measures. I know there's no data in orthopedics really at this point, but it is such a large effort from a PRO standpoint nationally, I'm wondering what people's perspective of where their role is in this work.

DR. BROWN: I'll take that one on. I think you mentioned Kahneman's book, *Thinking, Fast and Slow*, and there's an example in there about quality of life and dating. And you get a very different answer if you ask someone about their quality of life and ask them if they're in a relationship, in that order, versus if you ask them if they're in a relationship and then ask them about their quality of life.

And so my concern about the PROMIS approach is that if you

have a bank of questions, you can't validate a question, because every question you ask primes the responder for the next question, you know, so the order is very critical. So I don't know how validating a bank of questions works from a psychometric perspective. So in my approach, working with my healthcare organization, I am only taking validated instruments that are complete instruments. I actually think given that, the EQ-5D is wrong. I think EQ-5D should first give you the 0 to 100% scale where you rate your quality of life and then ask the five domain questions, because the way it is right now, your last question is depression. And then they say what's your quality of life from 0 to 100.

So my guess is that they unintentionally have some problems there, but that's the instrument that's validated, that's what we have, and that's what, you know, we use in our organization. But I fundamentally think that PROMIS, unless they, you know, take the questions and then put them in specific orders and validate them that way, it doesn't work as a validated instrument.

DR. TURKELSON: I think we need to be careful about the word validated. I'm not convinced that different investigators use that word in the same way, particularly in their own publications. Having had some occasion to look at this literature, I think that valid often means, okay, there is reasonable test, retest, or reliability. There's a non-statistical okay, it has construct validity. And perhaps it means a Cronbach's alpha. But it never

means, rarely if ever means that the questions were selected correctly, that the appropriate analyses were done to make sure that they were correct, that the weights of the questions are appropriate.

And one of the things I'm struck by is that I come from a world of comparative effectiveness research. And it would be interesting to see -- probably don't have the time for discussions here, but the role of the patient in determining what questions should be what. We all make assumptions.

And I think of a study performed by Chalmers in 2002 on patients with rheumatoid arthritis. And it's quite clear we all know that pain and function are what matters to patients with rheumatoid arthritis, so why bother to ask them? The answer is because they gave a different answer. The answer is that they were very, very concerned about fatigue, and nobody was even thinking about that.

So I'm wondering if when we get to -- if we shouldn't even consider that to be a valid instrument, it must have taken the patient perspective into account when the questions were framed. I'll get off my soapbox, but I'll leave it to you to comment on that.

DR. WYRWICH: I think that PROMIS measures hold great promise and that our government invested a lot into that development. I believe that they've also been tested in whether A is answered and B is answered versus B and A is answered, as well as the fact that the PROMIS domains developed so far both have the computer-adapted testing ability as

well as static forms that allow for them to be used.

The challenge for using them for regulatory is to go back and to take the items that are used and make sure that those are the items that would be fit for purpose in measuring in a specific disease area or a specific joint area, whatever the specific patient population is, because they were developed over a broad patient population. But once that validation can be completed, most of our major medical societies license at this time through computer-adaptive testing, and in fact, most individuals who take the GRE now, that's done through a computer-adaptive testing, which is very different than showing up at 8:00 in a large auditorium when I first did that.

DR. MIRZA: Great. If there's no other questions, then we'll conclude this session. And so thank you, everyone.

(Applause.)

DR. MIRZA: And before I introduce Dr. Mike Keith, I just have a couple of housekeeping issues to just make sure. First of all, Day 2 is in a different building. It's Building 66. And there's a different security entrance protocol for that. I believe everyone should have been provided all the instructions. And if not, if you're unsure of anything, the shuttles will definitely -- from the hotel will take you there, but just check in with the registration desk before you leave just to make sure you know where you're going because it's not here tomorrow morning. It's in Building 66 in the atrium.

The other thing is, if you haven't already, make sure you sign up for which breakout room you're going to be in, including Dr. Day, you need to sign up for your session tomorrow.

And the shuttles from here are leaving at 5:45 and 6:15, not 5 p.m. I think there was a misunderstanding that the shuttles were leaving at 5. It's 5:45 and 6:15.

Okay. And I now have the pleasure of introducing Dr. Michael Keith who is a past evidence-based medicine chair currently on the Appropriate Use Committee from Case Western.

DR. KEITH: Thank you all very much. This is going to be very quick. I was hoping to give a little bit of clinical perspective regarding both today's events and some of the, I guess, background problems that I foresee in getting to where we want to go.

I'll presume it's this way. Okay. I come from a background in which my research has been conducted with persons with spinal cord injury, guys paralyzed from the neck down, no bowel or bladder control, no sensation, and people who really do have significant clinical problems. And I've been working with NINDS for a long time developing implantable computers that replace the nervous system function that's lost. So, for example, these typically young men can regain the use of their hands, control of the bowel, the bladder, and stand and walk, which are pretty significant reversals of what is otherwise a pretty disabling condition.

What they would consider a substantially important clinical difference in spite of what we've been able to provide for them is a cure, that is, restoration to normal, replacement of all their neurologic losses. And some have, in good studies, said they would give up walking if that had to be a fair trade-off for getting everything else back.

And that's a substantially important clinical difference. And even our most -- Cohen's highest effect size doesn't record that, but there are people who have expectations that we're going to provide things like that to them. And so when we look at the minimal clinically important difference, sometimes it doesn't register. So I think the clinical group sometimes makes a big difference.

And a lot of the patients we've talked about today may have a hip problem or a knee problem, but today more and more of our people are living longer, getting more and more conditions in parallel, and taking a whole lot more drugs.

I know for having taken people with broken necks and giving them back this much neurologic function, that if you give them partial improvement, they know you can do better. And some of them will hold off on being treated, and some of them will still, once you give them what we consider pretty good for an MCID, will say, you know, you got my expectations right, but still, I know you can do better.

And that's one of the difficulties in looking at the evaluations

that clinicians receive from their patients and patients put down on our instruments is that every bit of improvement is only a partial improvement; nothing is ever perfect.

I will mention, too, that I think pain management is a day-to-day service as a right, entitlement, and as a product that's been skillfully inserted into our medical care organization, is now a way of life for many people. They wouldn't think about giving up their pain pills, their regular urine donations to prove that they're still taking their medicines correctly, because we have very good pain management today. We have very sensitive people and organized programs, and I think it's going to become harder and harder for us to look at the pain component of our outcome studies and say that we're going to do a washout and not have people really miserable from their back pain when we're evaluating their hands and hips because, in fact, they have had good pain management and they like it. A lot of them have implants for pain control.

I think we should give a lot more attention in the future as we develop our instruments to figuring out how to isolate the pain effect from the treatment that we're providing.

I find that a lot of patients are taking a lot of medications. Because we have electronic medical records in place, we are now seeing how many medicines they're taking every single time, including the med reconciliations on their over-the-counter meds; many people are taking 12, 15

meds a day in various patterns. And a lot of them interfere with each other, and a lot of them interfere with pain medicine, and a lot of them interfere with the drugs that we normally give people for pain medicine. I practically can't give tramadol to anyone anymore because so many people are taking other medicines, serotonin reuptake inhibitors, for example, that would result in an elevate seizure frequency.

So a lot of little meds that we could give are now not possible, and patients are migrating towards stronger meds. So their starting point for their definition of our pain description gets higher and higher.

I also wanted to comment on the very good discussion we had today about retrospective recall ratings. In one of our larger studies in which we did pre- and post-treatment assessments of a person's abilities, we videotaped each one of the sessions and interviews and then had them fill out the form so that at a year later when we asked them to do a recall comparison, for example, to recall what their pain score was before treatment, they got it wrong almost every time. Memory and recall are not to be trusted, and if we're designing or modifying our studies that ask for these comparisons rather than getting them prospectively or secondarily documenting them, I think we're going to make mistakes. We were pretty astonished that people were so well before their treatment that we probably wouldn't have treated them. But as you went back and looked at the videotapes, their ratings were much higher. So that's a caveat.

There are a lot of stakeholders. I mentioned these all because each one of these persons has a pretty big stake that I'll explain in a second for our device successes. I've taken devices through FDA previously and gotten them to market, so I know what a difficult process it is.

I want to mention from our quality assurance program at the American Academy of Orthopaedic Surgeons that there is a stepwise hierarchy of the activities that one needs to produce to have a good, quality program both as a hospital, a provider, and as a nation. And this is roughly the steps, although they have different words in different groups just like we were just talking about. But everything starts with a needs assessment, a clear definition of expectation, which in our clinical realm means education. The patient has to be trained and taught what they're going to get and how they're going to get it, and what decisions they have to make before surgery. You can't change expectations later by saying, oh, we want this surgery and things are going to be 10 times worse. That produces a very unhappy person.

When we provide technology such as the devices we're talking about, we give the patients choices as to the risks they're going to take and also the improvements that they might get or the costs that they might have to bear in order to get a better product or a newer product or something that isn't made of a metal they're allergic to.

These technology assessments are crucial to both the providers and the consumers making a correct decision about a device, and we don't

get enough of it. We certainly don't get the negative sides of it and sometimes until after we discover them in postmarket surveillance.

We know pretty much what we need to do with outcomes and how to make the measurements. That's a very elaborate and, I think, well-developed technology around the world today. We know how to write guidelines. We don't have the studies that we would like to have to support strong recommendations.

We're entering the world of accountability now because we have a framework for accountability that includes management at above the clinical practice level. We're now looking at accountability of many persons, not just institutions. We know that pay-for-performance medical care audits that we do are all designed to find the accountable physician and grade them. We know that hospitals have many ways of being graded, but on the issue of healthcare, safety, and efficacy, it's just now evolving. As Greg said, he belongs to an accountable care organization. Many of the other best organizations in the country haven't done that step yet.

We're looking at efficiency in value, and I think that, again, Greg's studies help exactly define how we ought to include that information in our future quality assessments.

And then I'm going to mention just the word quality at this stage because what I mean is performance measures by external organizations. There are quality organizations like NQF, and I think this

organization is evolving toward becoming a quality-determining organization as it writes better regulations that tell us what quality is built into device certifications.

If we take a look at the JCAHO description of accountability measures, I've been following their creation of something called ORYX, O-R-Y-X, a group of core measures that have been something like our practice guidelines and our other measures of quality. And they've recently changed them from quality measures to accountability measures. And they have good reason to do this because this extends the use of these measures to the healthcare organization they're auditing.

And I think that this next move toward accountability will affect all the organizations that are quasi-governmental and those that are truly governmental because we will want to assign accountability both to individuals, healthcare organizations, manufacturers.

Here's the manufacturers part. I believe every manufacturer in America right now is struggling with the questions of whether warranties will become part of the future of building devices in America. On the point of view of consumers, consumers are now seeing that they need contracts and eligibility for certain types of procedures at certain costs. I could easily foresee consumers being told, just like you get a 2-year contract on your cell phone, you'll get a five-year contract on your subscription or premium payments in order to get expensive healthcare because the healthcare

expense problem is out of control. Unless we have a good handle on everybody paying and everybody being insured, we can't really afford to provide the care and the devices that we're currently anticipating.

Providers are now seeing outcome-based reimbursement both as part of their HMO contracts and in a way an implied warranty on their day-to-day care that if they have bad outcomes, they're going to have bad reimbursement. They're going to be fined; if you don't fill out your electronic medical record, you're going to be fined or taxed. And these represent implied warranties on your behavior. As we understand them more closely in the next couple years, we'll understand how many deals we can make with which providers.

Hospitals now experience something called a warranty-on-never events. They bear the costs of never events occurring, and it's there to stay. CMS is not going to back down on the sort of things that are preventable and expensive.

I have mentioned manufacturers because we're now beginning to see, let's say, the companies that used to just go out of business and lose their exposure now wanting to stay in business and not being able to protect themselves from poor designs. I don't want to pick on any one company, but let's say a metal-type, metal-bearing design may be inferior to other designs, and therefore, they may have kind of an employed warranty there to provide a replacement device. We were experiencing that in our own city now. They

haven't quite gotten to the point of taking on greater warranty liability, and I know manufacturers don't want to see that.

Our current payers, of course, pay for everything, including the complications and including the costs for all of our patients getting better. But we're seeing changes in that as more and more of the cost of care is being passed on in the form of copayments and out-of-pocket expenses.

And then, finally, here we are as regulators, wondering how we're going to create the rules and regulation for enforcement of everyone else who's playing in this, let's say, postmarket device environment and trying to figure out that we need a registry in order to find the devices that aren't living up to their claims. So it's getting tougher and tougher not to have a tale, so to speak, on liability or the warranty of your products.

I see another thing shaping up now, too, and that is who's going to write the quality rules, and are we going to get everybody to subdivide the pie of regulation sufficiently as not to step on each other's toes. But we're currently seeing many organizations that are certifying either hospitals, providers, et cetera. And this race to regulate with very similar patterns of rulemaking and enforcement are everywhere now. The MOC for the providers here is one more example. The maintenance and certification of licensure in order to stay in practice is based on outcome measures. And there will be a time, I'm sure, when people decide not to do procedures that carry a chance of a permanent record of a failure or a bad result.

The FDA is just about to enter this with this enhanced postmarket surveillance being dependent on registries, being dependent on the perfect outcome measures. And we'll see how that goes.

One of the nice things I heard today, lesson learned, sample size should be powered for the MID -- except we just decided the MID was being replaced by the trial-specific responder paradigm. I have a verification that this is exactly correct. I heard it right; I wrote it down right.

UNIDENTIFIED SPEAKER: (Off microphone.) Wait, wait --

DR. KEITH: Please say it, please say it. I mean, in other words, what I've written up here is correct. Sample size should be powered for the MID. We're back in business.

Okay. And I don't want to drag this out. Just things for discussion tomorrow in our workshops. There's a lot of variability among our participants. I thought I was smart until I got into this room, and my god, I need to go back and take another fellowship at the FDA or something just to get caught up.

But as we work tomorrow, we need to work on our guidance document, at least as a first cut, to figure out how we're going to handle things like comorbidities, sex, not gender, gene expression and the genetic make-up of our patients -- thank you, Laura -- and then some functional equivalency of instruments to give the kind of guidance that came out in our questioning today; are there groups of things that kind of behave the same

way like, you know, all the short forms or all the healthcare things, so that our device makers can know that they're relatively safe to work with in these boundaries.

FDA could do a lot of favors here by, let's say, endorsing, maybe listing, do an ORYX-type project or assessment, so that we could use the established ratings of what are good and bad outcome measures that are already out there and have been evaluated several times.

And, again, on the pain issue, I don't believe we're going to be able to calibrate people. There is a discussion among my patients that say, Dr. Keith, you don't understand pain unless you've been in labor; you can't be in labor, you'll never understand pain. And then other guys who say, man, you've never been shot before; once you've been shot two or three times, you'll understand what pain is. These two kinds of perspectives help me calibrate pain. I don't understand pain yet.

Okay. Tomorrow we're going to solve all this. Remember that there's a 7:30 meeting tomorrow in the other building for all the moderators of the breakout groups. Have a pleasant evening, and thank you for your cooperation. It's been a wonderful meeting.

(Applause.)

(Whereupon, at 5:04 p.m., the meeting was adjourned, to reconvene the next day, November 28, 2012.)

C E R T I F I C A T E

This is to certify that the attached proceedings in the matter of:

GEORGIA TECH - TRIBES

MINIMALLY CLINICALLY IMPORTANT DIFFERENCE (MCID):

DEFINING OUTCOME METRICS FOR ORTHOPAEDIC DEVICES

November 27, 2012

Silver Spring, Maryland

were held as herein appears, and that this is the original transcription thereof for the files of the Food and Drug Administration, Center for Devices and Radiological Health.

_____

CATHY BELKA

Official Reporter